УДК 004.855.5 DOI: 10.14529/ctcr160302

ПРОБЛЕМЫ ФОРМИРОВАНИЯ ОБУЧАЮЩЕЙ ВЫБОРКИ В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ

И.Л. Кафтанников, А.В. Парасич

Южно-Уральский государственный университет, г. Челябинск

Правильное формирование обучающей выборки часто имеет решающее значение в задачах машинного обучения, что признаётся большинством специалистов в данной области. Зачастую решение задач машинного обучения сводится к грамотному формированию обучающей выборки. Несмотря на это, в современной литературе по машинному обучению вопросам формирования обучающей выборки почти не уделяется внимание, теоретическая база практически отсутствует. В настоящей статье постараемся исправить данный недостаток. В статье исследуются возможные проблемы и ошибки при формировании обучающей выборки, обобщается опыт авторов в решении задач машинного обучения, предлагаются теоретические модели для описания явлений, связанных с формированием множества обучающих данных, приводятся методы улучшения обучающей выборки. Даются практические рекомендации на основе разработанных теоретических моделей. В конце статьи представлены результаты экспериментов, демонстрирующие некоторые из проблем формирования обучающей выборки и методы их решения на примере задачи обучения деревьев решений.

Ключевые слова: машинное обучение, глубокие нейронные сети, деревья решений, обучающая выборка.

Формирование множества обучающих данных имеет принципиально важное значение для успешного решения задач машинного обучения. Часто задачи машинного обучения сводятся именно к правильному формированию обучающего множества. Ошибки в формировании обучающего множества обычно оказываются критичными и способны свести на нет эффективность самих алгоритмов обучения. Среди специалистов по машинному обучению общепризнанным считается, что наличие хороших обучающих данных гораздо важнее качества алгоритма обучения. В связи с активным развитием глубоких нейронных сетей в последнее десятилетие вопросы формирования множества обучающих данных принимают особенно важное значение, поскольку во многих задачах глубокие нейронные сети демонстрируют качество, существенно превосходящее остальные алгоритмы машинного обучения, однако, чтобы получить подобный выигрыш в качестве, необходимо использовать обучающее множество очень большого размера (до нескольких миллионов изображений, при этом обучение требует большого объема вычислительных ресурсов и может занимать несколько недель на многопроцессорном кластере), а также специальные методы расширения и имитации расширения обучающего множества, которые будут рассмотрены далее в статье. В то же время, в современной литературе по машинному обучению вопросам формирования обучающего множества уделяется недостаточное внимание, зачастую данные вопросы полностью игнорируются, недостаточно развита теоретическая база, объясняющая явления, возникающие в процессе формирования множества обучающих данных.

Введем некоторые определения.

Обучающее множество. Пусть имеется множество объектов X, множество допустимых ответов Y, и существует целевая функция $y^*: X \to Y$, значения которой $y_i = y^*(x_i)$ известны только для конечного подмножества объектов $\{x_1, \ldots, x_L\} \subset X$. Совокупность пар $X^L = (x_i, y_i)_{i=1}^L$ называется *обучающим множеством*. Задача обучения состоит в том, чтобы по обучающему множеству X^L восстановить зависимость y^*, то есть построить решающую функцию $a: X \to Y$, которая приближала бы целевую функцию $y^*(x)$, причем не только на объектах обучающего множества, но и на всем множестве X [1].

Метод обучения – это отображение $\mu: (X \times Y)^L \to A$, которое произвольному конечному обучающему множеству $X^L = (xi, yi)_{i=1}^L$ ставит в соответствие некоторую решающую функцию $a: X \to Y$. Также говорят, что метод обучения μ строит решающую функцию a по обучающему множеству X^L [1].

1. Модели множества данных

Вероятностная модель данных. Современное машинное обучение базируется на вероятностной модели данных. Считается, что обучающее множество $X^L = (xi, yi)_{i=1}^{L}$ является выборкой из генеральной совокупности некоторых объектов, при этом выборка должна отражать основные свойства генеральной совокупности. Также полагается, что вероятность появления объектов определенного типа в обучающей выборке равна вероятности появления данных объектов в генеральной совокупности. Однако эта модель имеет ряд принципиальных недостатков. Рассмотрим задачу распознавания позы человека. Допустим, имеется некоторый сценарий управления, состоящий из последовательности поз. В данном случае нас не интересует, какова вероятность появления каждой отдельной позы, нам нужно, чтобы любая из этих поз распознавалась также хорошо, как и все остальные позы в данном сценарии управления. Результаты в вероятностной модели сильно зависят от соотношения числа объектов разных типов в выборке. В задачах компьютерной безопасности система может иметь вероятность проникновения *0,001 %,* однако именно этот *0,001 %* будет активно искать злоумышленник, так что подобная оценка качества системы безопасности не имеет особого смысла.

В действительности, зачастую объекты в обучающую выборку выбираются далеко не случайно и независимо, вероятность появления объектов в обучающей выборке зависит от особенностей формирования выборки, которые следует учитывать при разработке систем машинного обучения. Например, известные базы изображений лиц *LFW* [2] и *FERET* [3] содержат изображения лиц совершенно разных людей, поскольку базы формировались разным образом. Алгоритм распознавания лиц, обученный на одной из этих баз, показывает плохие результаты при тестировании на другой базе. Естественно, разработка алгоритмов машинного обучения с использованием такого рода тестирования неэффективна. Особенно печально, что подобные явления могут происходить во многих случаях в менее явной форме. В данном случае свойства конкретной базы изображений не совпадают со свойствами генеральной совокупности, кроме того, иногда обучающие объекты не *«выбираются»* из реально существующих, а создаются искусственно (примеры будут рассмотрены далее). Таким образом, использование термина *«выборка»* не всегда уместно. Поэтому вместо словосочетания «*обучающая выборка»* в данной статье используется термин *«обучающее множество».* Кроме того, описание качества работы системы распознавания при помощи вероятностной модели малоинформативно – непонятно, на каких примерах система больше всего ошибается. Другими словами, вероятностную модель данных удобно использовать, только если данные представляют собой однородную (одномодальную) совокупность, что неверно в большинстве задач машинного обучения. Альтернативой вероятностной модели данных может послужить кейсовая модель.

Кейсовая модель данных. Все пространство возможных объектов X разбивается на некоторые виды данных (кейсы). Например, в задаче распознавания позы человека кейсом будет определенный вид позы (допустим, руки за головой). Каждый кейс характеризуется функцией принадлежности $f(x)$ (для каждого объекта определяет степень принадлежности данному кейсу) и важностью w. При этом важность кейса может меняться в процессе разработки системы в зависимости от требований заказчика. Упрощенно, кейс – это набор условий, которому удовлетворяет некоторая разновидность объектов из множества данных. Допускается произвольная вложенность кейсов. Внутри одного кейса можно использовать стандартную вероятностную модель данных. Заметим, что выбор кейсов в пространстве объектов может выполняться произвольно и служит исключительно для удобства описания работы системы. Здесь уместна аналогия со стратификацией выборки в статистике.

При использовании кейсовой модели при тестировании алгоритмов распознавания становятся более понятны общие свойства системы (в каких случаях она работает правильно или ошибается), кейсовая модель позволяет добиться устойчивости по отношению к изменениям относительного числа объектов разных видов в множестве. Также упрощается описание и понимание явлений при обучении алгоритма, так как в разных областях пространства объектов обучающее множество может иметь совершенно разные свойства (например, разную плотность данных), а вероятностная модель позволяет оценить только свойства обучающего множества или системы в целом.

2. Модель формирования наблюдаемых признаков

Обучающий объект x_i определяется набором наблюдаемых признаков $f_1, \ldots, f_q$. Будем считать, что каждый из этих признаков вычисляется некоторым неизвестным нам алгоритмом A_i расчета признака (индивидуальным для каждого признака) на основе значений некоторых переменных $z_1, \ldots, z_L$ (будем называть их генерирующие переменные). Эти переменные можно разделить на несколько видов.

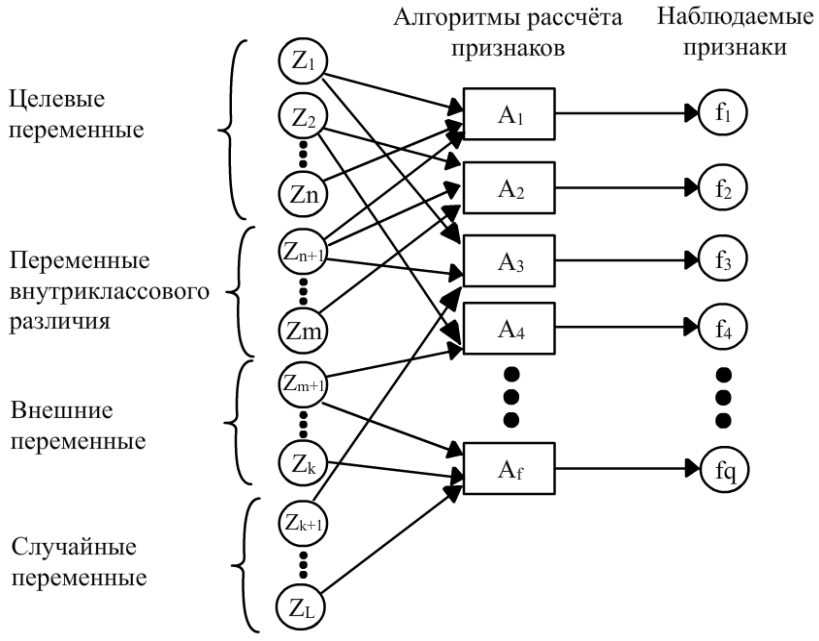
Целевые переменные. Переменные, определяющие внутреннюю структуру объекта и его принадлежность к тому или иному классу D_f. Знание значений целевых переменных гарантирует решение задачи распознавания.

Переменные внутриклассового различия. Переменные, определяющие внешние различия одинаковых с точки зрения задачи объектов (рост и цвет одежды пешехода в задаче поиска пешеходов). Знание значений данных переменных не обеспечивает решение задачи распознавания, однако может содержать ценную информацию для построения модели объекта. Не меняются с течением времени для одного и того же объекта. Обычно возможные значения переменных данного вида подчиняются некоторым ограничениям, которые могут быть учтены полезным образом. Заметим, что в зависимости от задачи одна и та же переменная может быть как целевой переменной, так и переменной внутриклассового различия. Например, пропорции лица человека будут являться целевыми переменными в задаче определения положения ключевых точек лица, и переменными внутриклассового различия в задаче поиска лиц на изображении.

Внешние переменные. Переменные, не зависящие от самого объекта, но влияющие на значения наблюдаемых признаков (поворот камеры, масштаб объекта, цвет фона). Могут произвольно меняться с течением времени для одного и того же объекта независимо друг от друга без особых ограничений. Знание значений данных переменных не дает полезной информации о самом объекте. Обычно в систему распознавания стремятся заложить инвариантность к изменениям таких переменных (например, дескрипторы изображений пытаются сделать инвариантными к повороту и переносу объекта интереса, изменению масштаба и аффинным преобразованиям).

Случайные переменные. Переменные, значения которых не зависят от природы объекта и воли человека. Обычно применяются для описания искажений и шумов на изображении. На основе их значений может быть задано некоторое распределение.

Будем называть объединение множеств целевых переменных и переменных внутриклассового различия внутренними переменными. На рисунке представлена получившаяся модель формирования наблюдаемых признаков.



Модель формирования наблюдаемых признаков

3. Способы генерации обучающего множества

Программная генерация. Генерация синтетических обучающих данных по некоторому алгоритму. В данном случае алгоритм генерации и его параметры определяют получающееся на выходе распределение в пространстве объектов. Целесообразно в процессе генерации варьировать как можно больше параметров. Однако если реально возможна лишь малая часть объектов из пространства параметров генерации или наиболее типичные объекты имеют сложную конфигурацию, метод оказывается неэффективным, так как сгенерирует много бессмысленных и мало действительно важных данных.

Сэмплирование. Призвано преодолеть недостатки предыдущего метода. Задается некоторое априорное распределение в пространстве объектов, и алгоритм пытается сгенерировать выборку из данного распределения. Применяются методы выборки с отклонением, сэмплирование по Гиббсу, схема Метрополиса – Гастингса, Marcov Chain Monte Carlo. Данные методы применяются для того, чтобы исследовать не все пространство объектов, а только его наиболее осмысленные части (обычно осмысленными является лишь очень небольшая часть потенциально возможных объектов, когда простой перебор объектов с проверкой их корректности может занять неоправданно много времени).

Закономерная модификация базового объекта. Имеется набор базовых объектов, обучающее множество получается путем непрерывной модификации их параметров (как правило, внешних переменных). Примеры – выборка кадров из видеопоследовательности, выборка поз человека из mocap-а. Данный метод применяется в тех случаях, когда нет возможности аналитически задать распределение возможных объектов в пространстве объектов, либо не подходит использование синтетических данных.

При использовании данного метода генерации обучающего множества устойчиво сохраняется большое число фоновых закономерностей. Кроме того, возникают проблемы, если полученное таким образом множество разбить в некотором отношении на обучающее и тестовое и использовать полученное тестовое множество для контроля за переобучением. Поскольку объекты исходного имеют сильную взаимозависимость, низкий уровень ошибки алгоритма на тестовой выборке не гарантирует отсутствия «заучивания» обучающих данных. При использовании метрических алгоритмов классификации самая выгодная стратегия в данном случае – всегда относить объект к тому же классу, к которому принадлежит ближайший обучающий объект. Пример ошибки подобного рода при использовании алгоритма *k-NN* приведен в работе [4] (при настройке гиперпараметров алгоритма методом скользящего контроля по множеству данных, полученному вышеописанным методом, всегда выбирался параметр $k = 1$). Поэтому при использовании данного метода формирования множества данных не следует применять метод скользящего контроля в чистом виде.

Выборка из базы объектов. Процессы формирования множества изображений с произвольными пейзажами или множества изображений лиц человека плохо описываются приведенными выше моделями. Мы не можем получить объект с любыми заданными внутренними параметрами, напротив, имеется фиксированный набор объектов с заранее определенными параметрами. В случае фотографий местности расположение этих объектов подчиняется строгим ограничениям. Все объекты можно разбить на группы, причем объекты внутри группы будут сильно похожи, а объекты из разных групп – различаться (можно сравнить изображения лиц преступников и изображения лиц политиков). Объекты, географически расположенные близко друг к другу, обычно более похожи, чем объекты, находящиеся далеко друг от друга (например, дома в одном и в разных городах). Если мы имеем дело с коллекцией фотографий, следует также принимать во внимание предпочтения оператора – люди гораздо чаще снимают красивые и необычные объекты (например, достопримечательности), чем однообразный лес или пустыню, разные люди делают снимки разных типов.

Проблема, возникающая при генерации обучающего множества данным способом – трудно гарантировать наличие всех принципиально важных типов объектов в множестве данных. Помочь в данном случае может переход к автоматической генерации или автоматическому изменению данных. Также следует производить выбор объектов из как можно более широкого набора групп объектов или мест съёмки (детектор лица, обученный на базе лиц преступников, возможно, будет не очень хорошо работать в общем случае).

Пример использования данной модели: если при решении задачи *hard samples mining* сложным негативным примером оказалось окно некоторого турецкого дома, то логично искать похожие изображения на снимках домов, сделанных в Турции (если имеется подобная информация об изображениях). Простая вероятностная модель формирования обучающего множества не позволяет проводить подобные рассуждения.

4. Возможные проблемы при формировании обучающего множества

Для оценки качества обучающего множества обычно используется его объем (количество обучающих примеров). Однако данная метрика не особо информативна. Во-первых, данных может быть очень много, но все они – одинаковые, во-вторых, даже если все объекты – разные, некоторые области пространства признаков могут остаться незаполненными, и, в-третьих, в самой процедуре формирования обучающего множества могут быть заложены ошибки. Рассмотрим некоторые возможные проблемы и ошибки при формировании обучающего множества.

Фоновые закономерности. В задачах машинного обучения объект может быть задан набором значений признаков $f_1, \ldots, f_q$ и значениями целевых переменных $z_1, \ldots, z_L$. Задача машинного обучения – найти закономерности между значениями наблюдаемых признаков и целевых переменных. При этом на основе каждого конкретного обучающего объекта, не принимая во внимание другие объекты, любую зависимость $y_k(x_i)$, характерную для данного объекта, можно посчитать всегда истинной. При рассмотрении большого числа разнообразных объектов из всех возможных закономерностей характерными останется лишь небольшое число действительно значимых закономерностей. Заметим, что на основании малого числа данных нет никакого способа отличить правильную закономерность от ложной. Будем называть подобные ложные закономерности, возникающие в результате нехватки данных, фоновыми закономерностями. По сути, некоторые виды переобучения заключаются в заучивании фоновых закономерностей. Пример фоновой закономерности – зависимость между классом изображения и цветом одного конкретного пикселя.

Отсутствие обучающих объектов определенного вида. Самый простой пример ошибки при формировании обучающего множества – если в нем отсутствуют данные определенного вида (не покрыта некоторая область пространства объектов, в кейсовой модели данных – отсутствуют объекты некоторого кейса), алгоритм не сможет правильно обучиться их классифицировать. При этом имеются в виду объекты в пространстве генерирующих переменных $z_1, \ldots, z_L$, а не в пространстве признаков $f_1, \ldots, f_q$.

Логично было бы добавить сюда и недостаточное количество обучающих объектов определенного вида, однако в разных случаях достаточным является разное число объектов, при разных алгоритмах обучения данная проблема будет проявляться совершенно по-разному, поэтому будем считать, что эта проблема входит в две следующие проблемы.

Отсутствие данных определенного вида относительно признаковой системы. Признаковая система $f_1, \ldots, f_q$ порождает некоторое разбиение множества данных на кейсы, каждому кейсу соответствует некоторый узкий набор значений признаков, при этом кейсов тем больше, чем более разнообразны и сложны признаки. Если некоторый из данных кейсов не будет покрыт объектами из обучающего множества или вероятностное распределение внутри кейса будет неверно отражать свойства генеральной совокупности, обучение может оказаться некорректным. Заметим, что при усложнении признаковой системы повышаются требования к обучающему множеству. Пример возникновения данной проблемы: в задаче поиска руки человека на изображении в качестве признаков используется гистограмма ориентированных градиентов с делением пространства поворотов на 64 ячейки, при этом в некоторые из ячеек не попадает ни одна рука из обучающего множества, и алгоритм может выучить, что рук с таким углом поворота не существует. При делении пространства поворотов на 32 ячейки проблема, возможно, исчезнет, а при выборе другой признаковой системы (например, интегральные изображения) данный недостаток обучающего множества вообще не будет проявляться при любых параметрах признаковой системы.

Некоторые из генерирующих переменных не варьируются. Важный частный случай проблемы отсутствия данных определённого вида. Очень часто при формировании обучающего множества часть генерирующих переменных имеет всегда одни и те же значения или очень узкий диапазон значений.

Разбалансировка. Неразумное с семантической точки зрения нарушение соотношений количества данных разного вида в рассматриваемом множестве данных, приводящее к необоснованному завышению влияния на результат одних и занижению влияния или полному игнорированию других данных, и, как следствие, к принятию неоптимальных решений. Самый простой пример – в обучающем множестве из-за особенностей его формирования примеров одного класса или типа гораздо больше, чем примеров другого класса или типа. Разные алгоритмы обучения имеют разную устойчивость к подобным проблемам. Разбалансировка особенно критична при использовании деревьев решений. Заметим, что разбалансировка – достаточно общий класс явлений, которые могут возникать не только в процессе формирования обучающего множества. Разбалансировка также возможна в задачах численной оптимизации и при тестировании алгоритмов распознавания (например, когда в задаче распознавания позы человека штраф за отклонение бедра от истинного положения превышает улучшения в точности определения остальных суставов, взятые вместе, хотя с практической точки зрения данная ошибка не имеет решающего значения). Часто проблема разбалансировки решается с помощью различных видов нормализации.

Внешние закономерности. В обучающем множестве $X^L = (xi, yi)_{i=1}^{L}$ могут присутствовать зависимости между внешними и целевыми переменными, которые алгоритм обучения может выучить как истинные из-за разбалансировок, особенностей обучения или признаковой системы. Например, в обучающем множестве $X^L = (xi, yi)_{i=1}^{L}$ все мужские лица сфотографированы в светлое время суток, а все женские – в темное. В таких условиях корректное обучение возможно лишь в том случае, если в признаковой системе $f_1, \ldots, f_q$ нет признаков, зависящих от освещенности изображения, как только такие признаки будут добавлены, алгоритм будет работать неправильно на выборке из другого источника. Наличие подобных закономерностей – пример ошибки при формировании обучающего множества. Признак наличия подобной проблемы – если после добавления более качественных признаков или признаков на основе информации другого вида алгоритм перестает корректно работать.

5. Способы добавления данных в обучающее множество

Добавление данных является одним из самых простых и эффективных способов улучшить качество обучающего множества. При этом простое добавление данных произвольного вида не всегда эффективно, часто требуется добавить данные определенной разновидности для повышения качества распознавания. Рассмотрим некоторые способы добавления данных.

Программная генерация. В случае использования синтетических обучающих данных удобнее всего сгенерировать недостающие обучающие примеры. Однако не во всех задачах допустимо использование программно сгенерированных данных. В таких случаях приходится применять более сложные методы добавления данных.

Data augmentation. Модификация имеющихся изображений с целью расширить обучающее множество. Активно применяется при обучении глубоких нейронных сетей, а также в условиях дефицита размеченных данных. Применяются сжатие/растяжение, горизонтальное отображение, поворот, случайный сдвиг в цветовом пространстве, случайное либо закономерное изменение некоторых пикселей. Считается, что добавление полностью случайного шума неэффективно, следует добавлять шум, обусловленный данными (только потенциально возможные в реальных данных искажения). Существенный недостаток данного метода – большинство фоновых закономерностей сохраняется.

Hard samples mining [5]. Классическая проблема в задачах поиска объектов на изображении – потребность в поддержании достаточного числа *hard negative samples* (обучающих примеров, которые похожи на объект интереса, но таковым не являются) в обучающем множестве. Сложность возникает из-за того, что в естественных условиях такие объекты встречаются редко, поэтому применяются специальные методы для их поиска и добавления в обучающее множество (*hard samples mining*). Ключевое предположение в данных методах – интересующие нас объекты сильно похожи между собой. Обычно применяются *data augmentation*, адаптивный поиск, поиск по шаблонам, методы на основе машинного обучения. Представляет интерес применение методов тематического моделирования для поиска сложных негативных примеров.

Имитация добавления данных. При обучении глубоких нейронных сетей обязательным считается применение метода *dropout* [6]: случайное обнуление активаций некоторых нейронов в

сети при подаче ей на вход очередного тренировочного изображения (обычно в каждом слое случайно выбирается 20–50 % нейронов). Без применения данной техники нейронная сеть «заучивает» большое количество фоновых закономерностей из-за того, что сложность модели превышает объем доступных данных. По сути, *dropout* – это имитация добавления данных в обучающее множество. Внутри алгоритма обучения мы имитируем изменчивость данных – на вход более глубоких уровней сети поступает случайным образом измененная версия реального изображения (подразумевается изображение, не вызвавшее бы активацию обнуленных нейронов), хотя таких данных на самом деле нет в обучающем множестве. Недостаток данного метода – может быть сымитировано добавление таких данных, которых в принципе не может быть в реальности, из-за чего может страдать точность распознавания. Представляет интерес создание модификаций данного метода, учитывающих природу данных.

Краудсорсинг. Поскольку для обучения глубоких нейронных сетей требуются огромные объемы вручную размеченных обучающих данных, для формирования обучающего множества активно используются сервисы краудсорсинга – пользователи сервиса за небольшую плату производят разметку «сырых» данных (например, указывают, какие объекты есть на данном изображении и где они расположены). Самый популярный из таких сервисов – *Amazon Mechanical Turk* [7]. Проблема данного метода – большое число ошибок в разметке, так как пользователи не всегда делают свою работу добросовестно, а иногда просто ошибаются, поэтому требуются специальные методы контроля ошибок в разметке. Обычно одну и ту же картинку дают разметить нескольким пользователям, а затем выбирают тот вариант разметки, который выбрало набольшее число пользователей. Также вводятся специальные метрики «добросовестности» пользователя.

6. Проблемы формирования тестового множества

При формировании тестового множества могут возникать те же самые проблемы, что и при формировании обучающего множества. Данные некоторого вида могут отсутствовать, из-за разбалансировки вклад объектов одного типа в финальную оценку качества может быть гораздо выше вклада объектов другого типа, наличие устойчивых внешних закономерностей не позволит корректно оценить работу алгоритма.

Будем называть источником данных запуск одного и того же алгоритма генерации множества данных с одними и теми же параметрами.

Принципиально возможны два типа тестирования алгоритмов обучения.

Обучение и тестирование на множестве данных из одного источника. Имеющееся множество данных $X^L = (xi, yi)_{i=1}^{L}$ разбивается на обучающую и тестовую выборку (обычно в отношении 70 и 30 %). Данный вид тестирования не позволяет выявить проблемы в формировании обучающего множества (поскольку проблемы, существующие в обучающем множестве, сохраняются и в тестовом множестве), поэтому он не может быть использован для оценки итогового качества системы в целом, однако может использоваться для оценки качества алгоритма обучения (например, для подбора параметров процедуры обучения глубоких нейронных сетей). Как отмечалось выше, данный метод нельзя применять, если множество данных получено путем закономерного изменения объекта (например, множество кадров из видеопоследовательности).

Обучение и тестирование на множествах данных из разных источников. При использовании данного вида тестирования решающий вклад в соотношение качества разных версий алгоритма может внести «заучивание» алгоритмом различий между обучающим и тестовым множеством (например, в обучающем множестве много лиц, повернутых вбок, но почти нет лиц, наклоненных вперед, а в тестовом множестве – обратная ситуация, в данном случае наиболее выгодно при обучении игнорировать те примеры, которых нет в тестовом множестве, и повысить важность тех, которых много в тестовом множестве). Поэтому слепое численное сравнение разных версий алгоритма в данном случае представляется малоосмысленным (точнее, является осмысленным только в том случае, если тестовое множество является целевым или тестовое множество заведомо лучше обучающего и не содержит ошибок формирования). Также имеет смысл вручную анализировать, на каких данных и из-за чего происходят ошибки, чтобы выявить возможные проблемы в формировании обучающего множества.

Как мы видим, ни один из представленных способов тестирования не является полностью правильным и не позволяет оценить итоговое качество системы. Здесь уместна аналогия с прин-

ципом неопределенности в квантовой механике. Нельзя одновременно измерить качество обучения и качество формирования обучающего множества. Поэтому в задачах машинного обучения тестирование качества следует вдумчиво проектировать, авторы не рекомендуют использовать «слепое» тестирование (то есть измерять качество алгоритма, не вникая в то, что именно измеряется и какие явления влияют на полученную оценку), подобный подход к тестированию может увести разработку в неправильном направлении.

7. Разбалансировка при обучении деревьев решений

Продемонстрируем важность правильного формирования обучающего множества на примере обучения деревьев решений (см. таблицу). В обучающем множестве имеется два принципиально различных типа объектов, причем по условиям эксперимента изначально в обучающем множестве присутствует разбалансировка и нехватка данных определенного вида (примеров первого типа в 5 раз больше, чем примеров второго типа, кроме того, примеров второго типа недостаточно). При этом имеется два тестовых множества, одно из них содержит только примеры первого типа, второе – только примеры второго типа. Деревья решений крайне чувствительны к проблемам разбалансировки.

Результаты обучения деревьев решений по несбалансированному и сбалансированному обучающим множествам

Число и вес обучающих примеров 1-го типа	Число и вес обучающих примеров 2-го типа	Качество классификации примеров 1-го типа на тестовой выборке	Качество классификации примеров 2-го типа на тестовой выборке	Общее качество классификации при соотношении вероятностей появления примеров 5:1 в пользу примеров 1-го типа	Общее качество классификации при равных вероятностях появления примеров обоих типов
500; 1	100; 1	82,48	49,2	76,93	65,84
500; 1	100; 3	81,52	57,31	77,48	69,41
500; 1	500; 1	78,88	80,95	79,22	79,91
500; 5	500; 1	83,68	66,18	80,76	74,93
500; 50	500; 1	83,8	20,48	73,24	52,14

При несбалансированном обучающем множестве примеры первого типа распознаются гораздо лучше примеров второго типа. Попробуем отдельно поднять вес примеров второго типа и добавить в обучающее множество примеров второго типа. Можно заметить, что оба приведенных выше способа позволяют заметно поднять качество распознавания примеров второго типа, при этом повышается даже совместное качество распознавания всего тестового множества, даже если предположить, что в реальных условиях количества примеров данных двух видов находятся в том же несбалансированном соотношении, что и изначально в обучающем множестве в нашем эксперименте. При этом добавление данных работает эффективнее, однако изменение весов (балансировка) технически удобнее, так как в случае ручной разметки данных не всегда есть возможность добавления новых данных, к тому же, экономятся машинные ресурсы (память и время обучения). Изначальная экспериментальная установка демонстрирует обучающее множество в естественном состоянии – присутствуют сильные количественные аномалии, кроме того, объектов некоторых типов не хватает. В последующих экспериментах показано заметное положительное влияние осмысленных точечных мер по улучшению обучающего множества на итоговое качество системы.

Выводы

Обобщен накопленный опыт по различным вопросам формирования обучающего множества, подробно проанализированы возможные проблемы при формировании обучающего множества, сформулированы модели, помогающие при описании явлений, происходящих при формировании обучающего множества (кейсовая модель данных, модель формирования наблюдаемых признаков), продемонстрирована практическая полезность этих моделей, составлены практические рекомендации по формированию обучающего множества в реальных задачах. Рассмотрены различные способы формирования обучающего множества, проанализированы их преимущества и не-

достатки. Рассмотренные модели будут полезны при анализе работы систем машинного обучения. Приводится список возможных ошибок при формировании обучающего множества, рассмотрены способы добавления данных в обучающее множество. На примере обучения деревьев решений показана важность правильного формирования обучающего множества.

Литература

1. Воронцов, К. Математические методы обучения по прецедентам (теория обучения машин) / К. Воронцов – http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf.

2. Labeled Faces in the Wild. – http://vis-www.cs.umass.edu/lfw/.

3. The Facial Recognition Technology (FERET) Database. – http://www.itl.nist.gov/iad/humanid/feret/feret_master.html.

4. Мангалова, Е. Прогнозирование мощности ветряных электростанций на основе непараметрического алгоритма k ближайших соседей / Е. Мангалова, И. Петрунькина // Доклады всероссийской научной конференции АИСТ'2013. – 2013 – С. 1–8.

5. Canavet, O. Efficient sample mining for object detection. / O. Canavet, F. Fleuret // Proceedings of the Asian Conference on Machine Learning (ACML). – 2014 – P. 48–63.

6. Dropout: A simple way to prevent neural networks from overfitting / N. Srivastava, G.E. Hinton, A. Krizhevsky et al. // The Journal of Machine Learning Research. – 2014 – Vol. 15, no. 1. – P. 1929–1958.

7. Amazon Mechanical Turk. – https://www.mturk.com/mturk/welcome.

Кафтанников Игорь Леопольдович, канд. техн. наук, доцент кафедры электронных вычислительных машин, Южно-Уральский государственный университет, г. Челябинск; kil7491@mail.ru.

Парасич Андрей Викторович, аспирант кафедры электронных вычислительных машин, Южно-Уральский государственный университет, г. Челябинск; parasichav@yandex.ru.

Поступила в редакцию 12 апреля 2016 г.

DOI: 10.14529/ctcr160302

PROBLEMS OF TRAINING SET'S FORMATION IN MACHINE LEARNING TASKS

***I.L. Kaftannikov**, kil7491@mail.ru,*
***A.V. Parasich**, parasichav@yandex.ru*
South Ural State University, Chelyabinsk, Russian Federation

Proper formation of the training set is often crucial in the problems of machine learning, that is recognized by most experts in machine learning, often solving machine learning problems is reduced to the competent formation of the training set. Despite this, in the modern literature on machine learning these issues given undeservedly little attention, although often it is the correct formation of the training set is crucial for solving practical problems, theoretical basis practically absent. This article is intended to correct this shortcoming. The article examines the potential problems and errors in the formation of a training set, summarizes the author's experience in solving machine learning tasks, offers a models for describing the phenomena, associated with the formation of a training set, methods for improving the training set are given. Practical recommendations, based on these theoretical models, are given. At the end of the article shows the experimental results demonstrating some of the problems of training set formation and methods for their solution by the example of learning a decision trees.

Keywords: machine learning, deep neural networks, decision trees, training set.

References

1. Vorontsov K. *Matematicheskie metody obucheniya po pretsedentam (teoriya obucheniya mashin)* [Mathematical Methods of Training on Precedents (the Theory of Machine Learning)] Avaliable at: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (accessed September 2015).

2. *Labeled Faces in the Wild.* Avaliable at: http://vis-www.cs.umass.edu/lfw/ (accessed September 2015).

3. *The Facial Recognition Technology (FERET) Database.* Avaliable at: http://www.itl.nist.gov/iad/humanid/feret/feret_master.html (accessed September 2015).

4. Mangalova E., Petrun'kina I. [Prediction Capacity of Wind Power Plants Based on Non-Parametric Algorithm, K Nearest to Neighbors]. *Doklady vserossiyskoy nauchnoy konferentsii AIST'2013* [Reports of the All-Russian Scientific Conference AIST'2013]. Ekaterinburg, 2013, pp. 1–8. (in Russ.)

5. Canavet O., Fleuret F. Efficient Sample Mining for Object Detection. *Proceedings of the Asian Conference on Machine Learning (ACML)*, 2014, pp. 48–63.

6. Srivastava N. Hinton G.E., Krizhevsky A., Sutskever I., Salakhutdinov R.R. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *The Journal of Machine Learning Research*, 2014, vol. 15, no. 1, pp. 1929–1958.

7. *Amazon Mechanical Turk.* Avaliable at: https://www.mturk.com/mturk/welcome (accessed September 2015).

Received 12 April 2016

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Кафтанников, И.Л. Проблемы формирования обучающей выборки в задачах машинного обучения / И.Л. Кафтанников, А.В. Парасич // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2016. – Т. 16, № 3. – С. 15–24. DOI: 10.14529/ctcr160302

FOR CITATION

Kaftannikov I.L., Parasich A.V. Problems of Training Set's Formation in Machine Learning Tasks. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2016, vol. 16, no. 3, pp. 15–24. (in Russ.) DOI: 10.14529/ctcr160302