Редкие сигналы учителя и активное обучение

Сергей А. Терехов

ООО "НейрОК Техсофт", г. Троицк.
http://www.neurokts.ru E-mail: alife@narod.ru

2011.11.15

Аннотация

Рассматривается проблема целенаправленного отбора обучающих примеров с известной информацией о классах. Предлагаются методы эффективного использования всех имеющихся данных, независимо от наличия меток классов. Развитые подходы применяются к задаче классификации дефектов труб магистральных нефтепроводов по данным магнитной внутритрубной дефектоскопии.

The problem of directed optimal selection of labeled training samples (Active Learning) is considered. Practical methods of utilization of all available data via Semi-Supervised Learning are proposed. The developed approach is applied to recognition and classification of pipeline defects from measurements of magnetic flux leakage defectoscopes.

Оглавление

Глава 1

Обучение на данных с редкими метками

1.1 Проблема

В практике разработки промышленных систем анализа данных типичной является ситуация, когда доступный корпус данных весьма обширен, однако основной объем наблюдений не содержит смысловых меток. В таких ситуациях традиционная постановка базовых задач *обучения машин* (классификации, регрессии, прогнозирования и др.) крайне затруднена.

Процесс получения новых меток классов включает обращение к экспертам в предметной области, что является крайне затратным, и может серьезно препятствовать успеху всего выполняемого проекта. В задачах регрессии и прогнозирования сбор выходных данных может основываться на новых экспериментах и измерениях, расходы на которые могут превысить доходы от внедрения разрабатываемой интеллектуальной системы.

Успешное решение должно содержательно ответить на три основных проблемных вопроса:

- *Пространство входных признаков.* Входные компоненты данных необходимо эффективно описывать безотносительно меток-выходов. Имеющейся информации о выходах заведомо недостаточно для адаптивного формирования входных признаков в процессе обучения.
- *Использование примеров без меток при обучении.* Это

предложение, на первый взгляд, может выглядеть парадоксальным - чем может помочь входной пример, если значение выхода для него не известно? В действительности, информация о распределении входов может значительно сужать класс вероятных моделей, что и может быть использовано, если сигналы учителя редки.

- *Стратегия сбора новых данных о выходах.* Направленный отбор данных для оценивания должен максимизировать информативность выходов ("сигналов учителя") при последующем обучении.

Первый из трех вопросов подробно отражен в литературе по методам обучения без учителя, на основе самоорганизации [1].

Обсуждению оставшихся двух вопросов посвящена данная лекция. Предметом нашего особого внимания будут пользоваться подходы, которые масштабируются до больших объемов данных [2] .

В формулах величины, относящиеся к многомерным данным, предполагаются векторными. Для указания на отдельные компоненты векторов используется необходимое число индексов. Для некоторых терминов и понятий в скобках сохранены их исходные названия на английском языке. Учитывая междисциплинарный характер конференции "Нейроинформатика", автор стремился минимизировать число вводимых математических понятий и формул.

1.2 Гибридное обучение с метками и без меток (semi-supervised learning)

Разумеется, отдельные примеры без меток, сами по себе, не могут повысить правдоподобие в задаче классификации с учителем. Однако, пространственное распределение множества обучающих примеров без меток может существенно изменять наши *априорные* представления о границах классов.

Рассмотрим простой пример. На плоскости имеются два обучающих примера, относящихся к разным классам. Априорное классифицирующее правило в этом случае дается прямой,

[1] Полное изложение вопроса можно найти в недавно переведенной на русский язык книге Т.Кохонена [1]

[2] В частности, вычисление матрицы всех попарных расстояний между примерами не доступно.

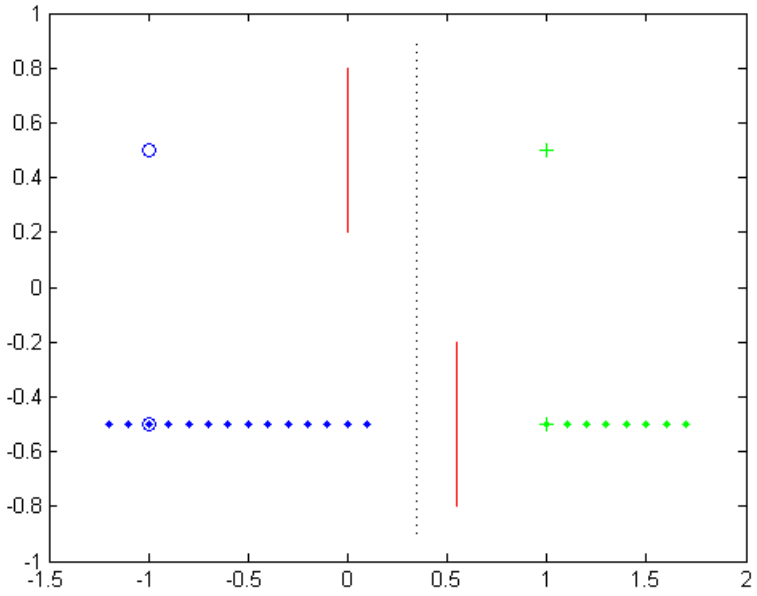

Рис. 1.1: Классификация двух примеров с метками (вверху) и при наличии дополнительных примеров без меток (внизу). Вероятная Байесова граница показана пунктирной линией.

равноудаленной от примеров (Рис 1, вверху). Это же правило максимизирует и правдоподобие в этой задаче классификации.

Пусть теперь дополнительно даны примеры без меток, расположенные, как в нижней части рисунка. Ясно, что теперь априорное положение классифицирующей прямой следует сместить. Полное решение задачи - классификатор, максимизирующий апостериорную вероятность по теореме Байеса, будет соответствовать некоторой промежуточной прямой, показанной на рисунке пунктиром. Примеры без меток *влияют* на вероятный ответ.

В приведенном примере выполнены все основные предположения о характере данных, которые благоприятствуют эффективному использованию примеров без меток в задачах классификации. А именно [2, 3, 4]:

- Данные образуют компактные кластерные структуры,
- Вероятные границы классов не пересекают областей, густо

населенных данными,

- Близкие в пространстве признаков примеры предпочтительно относятся к одному классу.

Значительное уклонение данных в прикладной задаче от таких предположений может приводить к ухудшению точности классификации при учете примеров без меток [3].

1.2.1 Алгоритмы классификации на графах

Одним из популярных прямых путей, непосредственно использующих высказанные предположения о данных в задаче классификации, являются методы, основанные на графах. Граф, узлами которого служат все обучающие примеры (и с метками, и без), а ребрами соединены пары "близких" примеров, позволяет выделить однородные области, заполненные данными. Строится классификатор с границами классов, пересекающими минимальное число ребер графа.

Рассмотрим формальную графовую структуру, в которой узлам сопоставлены все из имеющихся N входных векторов данных. Назовем L множество индексов примеров с метками, и будем использовать U для примеров без меток. Ограничимся для простоты задачей классификации входных данных в два класса. Сопоставим узлам графа функцию f, имеющую смысл вероятности класса 1. Для узлов с известными метками эта функция принимает значения 0 или 1. Для оставшихся узлов значения функции из интервала [0..1] подлежат определению.

Наша цель - придерживаясь высказанных предположений о характере данных, получить условия для нахождения f для всех примеров без меток.

Для описания соседства данных, которое лежит в основе возможных кластерных структур, введем в рассмотрение ребра между вершинами графа. Степень соседства для пары узлов j и k описывается весовым коэффициентом w_{jk}:

$$w_{jk} = exp(-\beta \|x_j - x_k\|) \tag{1.1}$$

Ребра с коэффициентами меньше некоторого порога w_0 удаляются, что в общем случае приводит к разреженной структуре с числом ребер, пропорциональным числу узлов N. Скейлинг не сложнее $O(N)$

[3] Читатель без труда построит соответствующие контрпримеры.

является абсолютно необходимым условием для того, чтобы алгоритм мог претендовать на практическую полезность[4].

Ребра, инцидентные каждому узлу, задают его топологическую окрестность. По предположению о схожести классов для схожих обучающих примеров (здесь - узлов одной окрестности) функция-классификатор удовлетворяет условию:

$$f_j = \frac{\sum_k w_{jk} f_k}{\sum_k w_{jk}}, \quad k \in (L+U) \tag{1.2}$$

для всех примеров j. Это – одна из возможных форм данного условия. Она приводит к семейству так называемых *гармонических* функций, удобных для использования на практике. Формула (1.2) автоматически гарантирует быстрые изменения функции (в областях границ между классами) только между узлами, не связанными ребрами, либо при слабой степени связности w.

Гармоническое условие дополняется ограничениями на известные значения для примеров с метками

$$f_j = y_j \in \{0, 1\}, \quad j \in L \tag{1.3}$$

Если в известных метках y_j допускаются некоторые ошибки, может использоваться слабая форма

$$\|f_j - y_j\| \to 0, \quad j \in L \tag{1.4}$$

Заметим, что соотношения (1.2) - (1.3) представляют собой систему линейных алгебраических уравнений(СЛАУ), которые могут быть совместно решены множеством хорошо разработанных методик. Методы для СЛАУ выходят за рамки данной лекции (см. например, [5]).

Граф и коэффициенты определены достаточно абстрактно и могут быть заданы различными способами. Одна из популярных альтернатив - использование в качестве инцидентных ребер множества из заданного числа пространственно ближайших соседей для каждого узла. При этом значения ненулевых элементов матрицы w могут быть выбраны постоянными.

[4]Современная литература содержит достаточное число описаний алгоритмов с базовой сложностью N^2 и старше. Поскольку (итерационное или прямое) решение полученных уравнений часто добавляет еще одну степень N, то полученные построения при типовом промышленном объеме данных $N \sim 10^6$ ориентированы, видимо, на компьютеры следующих поколений.

Для иллюстраций система (1.2) - (1.3) может быть приближенно решена методами простых итераций (Зейделя). Для этого уравнение (1.2) записывается в форме ращепления по итерациям, и на каждом шаге обновления вычисляются для случайно выбранного узла[5].

$$f_j^{(n)} \longleftarrow \frac{\sum_k w_{jk} f_k^{(n-1)}}{\sum_k w_{jk}}, \quad k \in (L+U) \tag{1.5}$$

Заслуживают внимания также методы Монте-Карло, связанные со случайными блужданиями на графе. Пусть из каждой вершины стартует траектория, в которой вероятности переходов к соседним узлам по ребрам графа пропорциональны весовым коэффициентам этих ребер. Траектория заканчивается в одном из множества узлов, для которых значение метки класса известно. Тогда в качестве оценки вероятности принадлежности данного узла каждому из классов выступает статистическая частота меток классов в достигнутых конечных точках траекторий. Заметим, что каждая разыгранная траектория обновляет частоты классов не только начальной точки, но и всех промежуточных точек на траектории. В итоге, дисперсия оценок быстро уменьшается для узлов, расположенных в окрестности узлов с известными метками. Достоинством метода Монте-Карло является его естественная параллельность.

Постановка задачи в форме (1.5) привлекает нас здесь потому, что вычисления могут быть эффективно реализованы на нейроподобных аппаратных структурах. Это является важным фактором в некоторых специальных приложениях. В целом же, графовые модели, несмотря на их простоту и теоретическую ясность, обладают рядом недостатков, сдерживающих их широкое применение. А именно:

- Графовые алгоритмы дают лишь *трансдуктивное* обучение. Другими словами, значения функции-классификатора определены только в точках имеющихся данных. Все интересующие примеры (и с метками, и без) должны быть заранее определены. При добавлении новых данных необходимо заново перестроить классификатор.

- В приложениях, в условиях шума в данных, весьма непросто построить хороший граф соседства. Проблема особенно обостряется, если в разных областях пространства данных

[5]Разумеется, такой "алгоритм"не может использоваться на практике для числа примеров, превышающих несколько сотен.

ближайшие примеры отстоят друг от друга на разных масштабах. Эта проблема не так остра для графов на основе ближайших соседей, однако в этом случае фиксирование числа соседей может нарушать естественную кластерную структуру в малозаселенных областях данных. Построение сбалансированного графа в общем случае представляет собой нерешенную задачу[6].

- Графовые модели не масштабируются при росте объема данных. Надежные вычисления (и, прежде всего, построение самого графа) при объеме данных выше нескольких десятков тысяч примеров затруднены.

Перспективными для графовых методов являются приложения, в которых структура графа определена и задана изначально. Примером такой задачи, актуальной в последнее время, служат графы социальных сетей. Эта тематика весьма обширна и заслуживает отдельной лекции.

Здесь мы обратимся к методам обучения на примерах с редкими метками, в которых плотность данных моделируется и используется локально (поточечно).

1.2.2 Нейросетевой классификатор для данных с метками и без меток

Нейронная сеть может быть традиционными способами легко обучена только на порции примеров с метками. Однако, если доля примеров с метками заметно мала, то возникают, по меньшей мере, два вопроса:

- Трудно надеяться на хорошие обобщающие свойства такой модели - нейронная сеть не в состоянии сформировать существенные для задачи категории в выходных слоях.

- Еще б'ольшие проблемы возникают на входах. При полноценном обучении (на достаточной выборке данных) входные нейроны, в теории, автоматически формируют набор информативных признаков (или "черт"). Даже при богатых данных этот процесс, в действительности, достаточно сложен, и полученные признаки могут оказаться далекими от оптимальных. На эту проблему специально обращают внимание в последнее время в связи с

[6] В действительности, для заданного графа можно указать относительно небольшое число узлов, изменение которых нарушает структуру связности данных. Эта задача неустойчива.

методами глубокого автокодирования (deep autoencoding, [6]). Использование только информации на входах для формирования входных признаков неэффективно.

В нашем случае редких меток для формирования содержательных входных признаков имеющейся информации о метках *заведомо недостаточно*.

Здесь предлагается простая и ясная идея - использовать для решения задачи классификации готовый базис входных слоев, который формируется при адаптации нейронной сети к пространственному распределению (плотности) данных. Для оценки плотности могут быть использованы все имеющиеся примеры (в том числе, без меток).

Разумеется, информативный базис для задачи оценивания условной вероятности $p(y|x)$ трудно построить на основе апроксимации только $p(x)$, однако в последнем случае будет рационально описана топология входного пространства. Примеры, относящиеся к одной связной области, будут порождать схожую активность в выходных нейронных слоях. Тогда формально достаточно одного примера с меткой, чтобы правильно классифицировать всю однородную область. Также автоматически гарантируется и отсутствие рассечения плотной однородной области границей классов - в построенном по $p(x)$ базисе просто нет соответствующих элементов.

Общая задача аппроксимации плотности уже рассматривалась автором в одной из предыдущих лекций [8]. Нейросетевое решение состоит в отделении истинных примеров от случайных точек с известным (например, пространственно постоянным) распределением. Если при обучении истинным примерам присваивается метка $+1$, а случайным -1, то выход $a(x)$ такого классификатора дает оценку локального значения плотности истинных данных

$$\rho(x) \sim \rho_0 \frac{1 + a(x)}{1 - a(x)} \tag{1.6}$$

где ρ_0 - плотность фоновых случайных точек.

Для решения исходной задачи классификации в нейронную сеть, аппроксимирующую плотность, добавляется еще один выход, кодирующий класс примера[7].

[7] Одного дополнительного нейрона достаточно для случая двух классов, с метками ± 1. Для случая $k > 2$ классов может использоваться k выходных нейронов. Такие

Обучение нейросети с двумя выходами проводится по следующему алгоритму (с учителем). Веса первого выхода, ответственного за моделирование плотности, обучаются всегда. Веса второго нейрона, занятого исходной задачей классификации, используются для формирования градиента только если пример снабжен известной меткой класса. После каждого примера из входного потока для обучения предъявляется пример со случайным входом (без метки).

Таким образом, первый выход нейросети интенсивно участвует в обучении, оценивая плотность истинных на фоне случайных примеров. В скрытых слоях нейронов происходит формирование базиса входных признаков. Второй выход использует текущее состояние этого базиса для классификации в соответствии с метками классов. Этот выход практически не влияет на базис, т.к. сигналы учителя предполагаются редкими.

Заметим, что если задача решается в обычной постановке, когда метки всех обучающих примеров известны, предлагаемая модель ведет себя, как обычная нейросеть. Попутная аппроксимация плотности выполняет роль специфической регуляризации[8].

Аппроксиматор плотности играет самостоятельную практическую роль в задаче классификации. А именно, его выход для новых примеров указывает на степень их принадлежности к исходной обучающей совокупности данных. Таким образом, мы получаем не только сам результат классификации, но и количественную меру его надежности. Это неоценимо в приложениях, связанных с оценками риска при прогнозировании. Платой за такие удобства служит удвоение набора обучающих данных и несколько более сложная архитектура нейронной сети.

Для иллюстрации рассмотрим модельную задачу классификации точек на плоскости в предельно сложных условиях, когда из 2,000 примеров метки заданы лишь для двух из них. Обучение без учета примеров без меток приводит к классификатору, имеющему мало общего с ожидаемым результатом (вертикальная линия между двумя точками

модели обычно не рекомендуются при традиционном обучении с учителем (выходные нейроны становятся искусственно зависимыми), однако в случае с редкими метками рационально использовать один общий базис для всех классов.

[8]Действительно, класс функций, описывающих решающие правила классификатора *при условии* одновременной аппроксимации плотности примеров, *сужен* в сравнении с классом функций без этого условия. Таким образом, в соответствии с теорией А.Н.Тихонова, попутная аппроксимация плотности является *регуляризатором* для исходной задачи.

с метками). Применение попутного обучения плотности дает близкий к идеальному ответ - различные пространственно связные области данных относятся к разным классам (Рис. 2). Для полноты картины на Рис 2 также изображена результирующая аппроксимация плотности, выделяющая области, в которых классификация надежна.

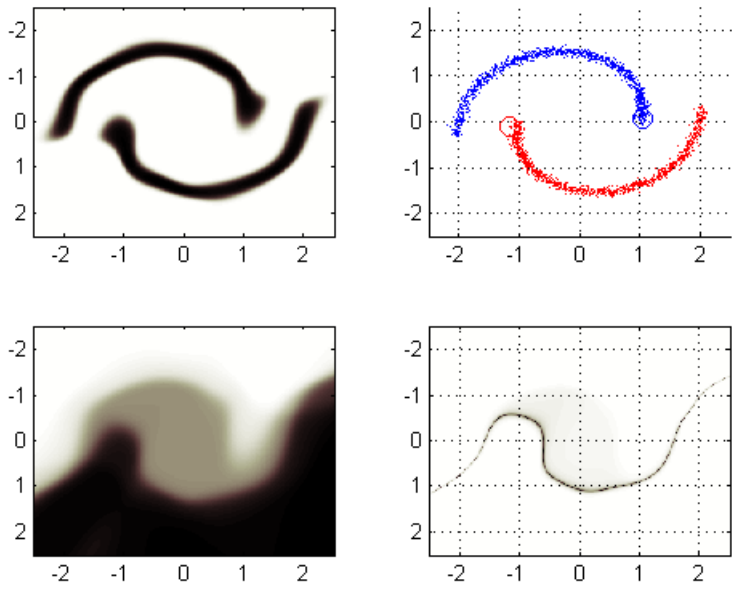

Рис. 1.2: Задача классификации 2,000 примеров без меток и двумя примерами с метками. Слева вверху - аппроксимация плотности, слева внизу - распределение вероятности класса 1, справа вверху - точки двух классов, кружками выделена пара примеров с метками. Классы для всех примеров определены без ошибок. Справа внизу - оценка границы классов (линия уровня 0 для выхода нейросетевого классификатора).

В завершение этого раздела заметим, что для формирования входных признаков можно было бы использовать и оригинальные модели deep autoencoding [6, 7]. Однако они связаны, так или иначе, с полным восстановлением входа на выходе, если нейронная сеть "узнала" данный входной образ. Такое восстановление гораздо более затратно, чем формирование скалярного сигнала "узнавания" от аппроксиматора

плотности, который достаточен[9] для подкрепления при обучении распознаванию и классификации.

1.3 Активный отбор данных при обучении (active learning)

В предыдущей главе был предложен практический инструмент, позволяющий эффективно использовать при обучении всю имеющуюся информацию о данных. Обратимся теперь к вопросу о сборе новой информации. При этом, по-прежнему, будем предполагать, что нет дефицита в данных с неизвестной принадлежностью к классам. Поэтому сконцентрируемся только на проблеме сбора новых меток для уже имеющихся примеров.

1.3.1 Общая задача планирования сбора информации

Нашей целью является такой процесс получения новых меток, который скорейшим образом способствует уменьшению ошибки последовательно обучающегося на новой информации классификатора.

Эта задача относится к области оптимального планирования, которая допускает несколько базовых постановок, в зависимости от горизонта плана и характера неопределенности в ценности будущих испытаний.

Для пространств невысоких размерностей могут использоваться определенные фиксированные планы эксперимента, например, латинские гиперкубы, с выбором для каждой экспериментальной точки плана ближайшего к ней примера из имеющегося набора данных.

Если зафиксирован полный ресурс (бюджет) на экспериментирование, то задача может быть сформулирована, как задача стохастической оптимизации [9]. Это – отдельный широкий класс прикладных методов, изложение которого уведет нас от тематики лекции.

[9]Это, разумеется, отражает личную точку зрения автора. Имеются определенные биологические свидетельства того, что живые системы занимаются восстановлением входов. Возможно, этот механизм более устойчив и допускает локальное обучение. В предлагаемом подходе сигнал аппроксиматора плотности распространяется в обратном направлении до входов, что до определенной степени глобализует обучение. С другой стороны, возможные биологические механизмы, отвечающие представлению о плотности, гораздо сложнее обнаружить экспериментально, так как они связаны с малым числом отдельных подкрепляющих нейронов.

В случае, когда информационные вклады отдельных наблюдений (здесь - меток классов) независимы[10] или имеют определенную корреляционную структуру, задача сводится к известной задаче о многоруком бандите (см. популярное изложение [10]).

В целом, исчерпывающего теоретического решения задача планирования в общем случае не имеет. Однако в нашем частном случае, во-первых, имеется надежный инструмент измерения качества уже собранных меток (а, именно, обученный на них классификатор), а во-вторых, имеется попутная информация о характере уже имеющейся модели в применении к новым, перспективным для сбора меток, точкам. Это позволяет провести оценивание и ранжирование всех примеров данных на каждом шаге добавления меток, приводя к эффективным практическим стратегиям.

1.3.2 Ранжирование и активный выбор примеров

Рассмотрим задачу выбора информативных примеров в он-лайн постановке (с неограниченным горизонтом плана). На каждом шаге выполняется следующая последовательность действий:

- Производится обучение классификатора на всей имеющейся[11] совокупности данных с метками и без меток. Обучение может и не использовать предложенный в этой работе подход с попутной аппроксимацией плотности, однако его применение дает ощутимые практические преимущества. В частности, обучение может быть продолжено до полного отсутствия ошибок на примерах с метками, что дает ясный критерий останова.

- Построенная модель применяется (в режиме распознавания) ко всем данным без меток. Результаты классификации оцениваются по одному из выбранных информационных критериев (о конкретных критериях – ниже) и сортируются в порядке убывания оценки. При применении попутного обучения плотности, из

[10] В исходном виде, информация, доставляемая каждой меткой, конечно, зависит, от того, какие еще метки известны. Эффекты этой зависимости могут быть ослаблены переходом к кластерной модели входных данных, с рассмотрением испытаний каждого кластера, как целостной сущности. Каждый кластер описывается ровно одним примером-представителем, метка которого запрашивается в эксперименте. Тогда, при условии пространственной разрешимости кластеров, такие испытания могут считаться независимыми.

[11] При большом дефиците меток проблема формирования тестовых выборок и, вообще, подходов к тестированию, должна рассматриваться отдельно.

упорядоченного списка удаляются примеры, индикатор локальной плотности для которых ниже выбранного порогового значения. Это позволяет избежать малополезных оценок меток для выбросов и слабо заселенных областей в данных.

- Пример (или выбранное число примеров) с наивысшими рейтингами предлагается эксперту для назначения метки класса (либо в условиях, описываемых данным примером, производится эксперимент, в результате которого определяется класс). Если анализ выбранного примера по каким-то причинам затруднен, может использоваться один из следующих по рейтингу примеров.

- Принимается решение о пригодности текущей модели (например, проводятся ее промышленные или лабораторные испытания), либо о продолжении сбора данных для уточнения модели. В последнем случае процесс повторяется.

Такой логике, в основном, следуют весьма актуальные в последнее время сценарии интерактивного обучения машин и человеко-машинных систем (interactive machine learning, см. [11]). Непосредственная вовлеченность эксперта или пользователя в процесс синтеза интеллектуальной модели может стать решающим фактором успеха прикладного проекта, и значительно повышает уровень доверия к искусственной системе (хотя она продолжает оставаться "черным ящиком").

Основная задача рекомендательного алгоритма отбора данных сводится, *формально*, к уменьшению длины последовательности обработанных экспертом примеров, и, *неформально*, к нахождению "интересных" примеров, отражающих разносторонние аспекты исследуемой системы или процесса.

Остановимся теперь на количественных критериях, которые используются при отборе данных.

Основной критерий, который интенсивно эксплуатируется, например, в моделях машин опорных векторов - это степень близости примера к текущей оцениваемой границе классов. Его числовым выражением служит энтропия прогнозируемого распределения вероятностей классов в данной точке. Примеры с высокой энтропией выходов, если их метки станут известными, окажут максимальное влияние на положение границы.

В работе [12] предложено интересное обобщение этого критерия для случая нескольких классов. Вместо вычисления энтропии, которая при

большом числе классов слабо различает наличие или отсутствие *одного* класса-лидера, в качестве критерия предложено использовать разницу между вероятностью класса победителя и следующего за ним наиболее вероятного класса.

Другие (эвристические) критерии могут использовать энтропию активностей в слоях скрытых нейронов. При эффективном узнавании знакомого примера, ему обычно соответствует всплеск активности относительно малого числа нейронов в скрытых слоях. Напротив, хаотическая активность большого числа нейронов может возникать, если пример "новый", и для него еще не сформировался компактный код.

В реализациях алгоритмов активного обучения в ООО "НейрОК Техсофт" этот критерий используется, как второстепенный, для снятия неопределенности при выборе примеров в случае слабого контраста энтропийных критериев.

При обучении с одновременной оценкой матожидания и дисперсии выходов (пример – ранее описанная автором в лекциях система CNet [13]), для предпочтительного назначения меток могут рекомендоваться примеры с высокой дисперсией (неопределенностью) выходов. Этот критерий идентичен подходу, применяемому в алгоритмах на основе условных гауссовых процессов и опорных векторов (Relevance Vector Machine и Informative Vector Machine [14]).

Могут использоваться и иные, например, геометрические, принципы рейтингования, как то удаленность примера от других примеров с известными метками, либо выбор примера, способного изменить класс у максимального числа примеров, если их классифицировать методом ближайшего соседа.

Применяемые на практике стратегии используют комбинации формальных и неформальных оценок, а также специфику задачи. Специфика может состоять в указании априорной функции полезности (utility function), определенной для входных (известных) компонент данных. В приложениях такие критерии могут оказаться особенно успешными.

Подводя итоги этого раздела, продемонстрируем свойства различных подходов к активному выбору меток на сложной модельной задаче классификации трех вложенных спиралей на плоскости. Это задача классификации в три класса. Она характеризуется экстремально сложной геометрией решающих правил. При зашумлении данных классы также могут частично перекрываться.

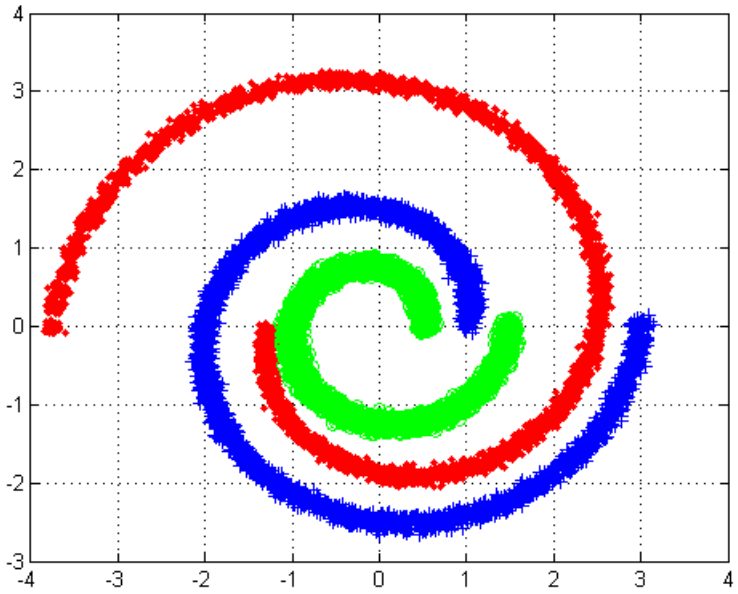

Рис. 1.3: Данные в модельной задаче классификации в три класса (15,000 примеров).

Моделирование процесса активного выбора примеров для всех проведенных расчетов начинается с состояния с тремя известными метками (по одному примеру для каждого класса). Сравнивались 4 стратегии выбора примеров для добавления в обучающую выборку. Три из них не используют примеры без меток: случайный выбор, выбор примера с минимальной энтропией выходов, выбор примера с минимальным опережением выхода-победителя над следующим по уровню активности выходом. В 4-м варианте производился выбор примера с минимальной энтропией, но при обучении использовалась попутная аппроксимация плотности, с учетом всех примеров без меток.

Результаты эксперимента типичны для данных методов. Стратегии, не использующие примеры без меток, показывают высокую эффективность в сравнении со случайным выбором. При случайном выборе не удается достигнуть высоких уровней точности, так как для исправления ошибок текущей классификации необходим выбор примеров из областей тесного контакта классов. Однако случайное

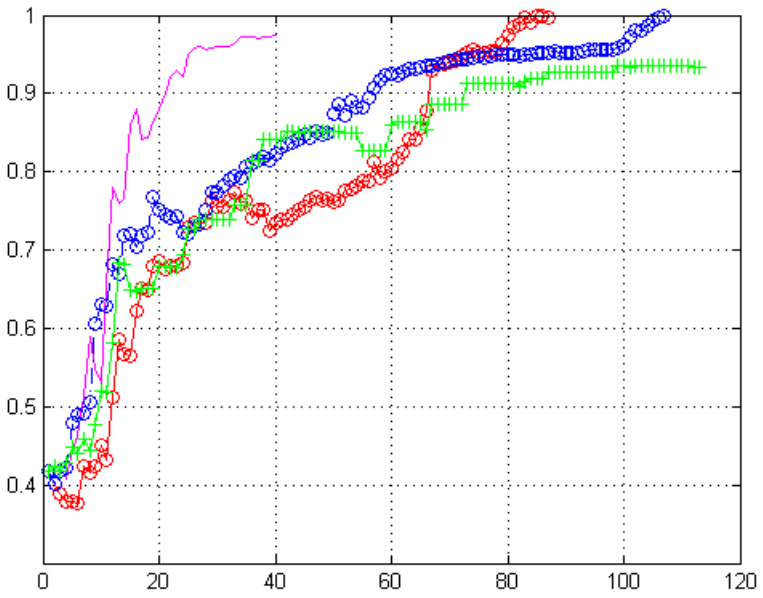

Рис. 1.4: Точность решения задачи классификации при последовательном добавлении обучающих примеров с метками. Стратегии активного обучения: отбор по максимуму энтропии выходов (кружки, сплошная линия), опережение первого победителя над вторым (кружки, штрих-пунктир), случайный выбор (плюсы, сплошная линия) и отбор по энтропии с попутной аппроксимацией плотности (сплошная линия без маркеров). Случайный выбор не позволяет уменьшить ошибку в областях "слипания" данных из разных классов.

попадание в эти области относительно маловероятно.

Попутная аппроксимация плотности кардинально улучшает результат. Точность в 95% достигается при числе примеров 23, в то время, как для других методов для сравнимых уровней точности необходимо более 70 примеров с метками. Геометрия области, заселенной данными, играет решающую роль при классификации.

Глава 2

Анализ дефектоскопических данных

2.1 Практическая задача: классификация дефектов промышленного трубопровода

Описанные в предыдущих разделах подходы применялись к промышленным задачам классификации дефектов стенок промышленных нефтепроводов. Информационным сигналом являются локальные искажения потока постоянного магнитного поля, насыщающего стенку трубы. Измерения проводятся промышленными внутритрубными дефектоскопами, следующими в потоке нефти, промагничивающими стенки и измеряющими сигналы утечки потока. В специальной литературе этот метод диагностики носит название "метод утечки магнитного потока" (MFL - magnetic flux leakage, [15]).

Собственно этапу классификации предшествовали технологические этапы обнаружения и распознавания областей дефектов на фоне сигналов от конструкционных элементов трубопровода и источников "шума" вследствие неидеальности трубы и самих диагностических приборов. Затем, для выявленных областей аномалий магнитного сигнала, проводилось их количественное описание на основе системы информативных признаков (выбор которых представляет собой отдельную сложную задачу).

При использовании методов магнитной дефектоскопии основная проблема состоит в установлении связи двумерного магнитного отклика вблизи плоскости расположения дефекта, например у стенки трубы, с истинными трехмерными параметрами дефекта, такого как коррозийная

Рис. 2.1: Измерительный прибор - внутритрубный магнитный дефектоскоп (http://www.ppsa-online.com).

потеря металла, механическая риска и др. Задача получения 3D информации из 2D сигнала является некорректно поставленной, и ее устойчивое решение в данном частном случае может опираться на априорную статистику характерных сигналов, параметры дефектов для которых известны.

Каждый из выбранных для анализа 647 дефектов был представлен 24 признаками, включающими свойства магнитной системы, параметры формы магнитного сигнала, контекст расположения и другие свойства. Глубины проникновения в стенку для этих дефектов считались известными[1].

В контексте данной лекции, задача представлена в виде классификации векторов признаков в семейство из трех классов, отвечающих различным глубинам дефектов. Эта подзадача важна как для разработки последующей нелинейной регрессионной модели, прогнозирующей размеры дефектов, так и для других задач распознавания (например, классификации дефектов по их типам).

Обучение нейросетевого классификатора с двумя скрытыми слоями нейронов[2] проводилось по методике, описанной выше в лекции, с попутной аппроксимацией плотности. Использовалась Байесова

[1]Размеры различных дефектов известны с разной степенью точности. Часть дефектов была нанесена на стенку трубы искусственно, в полигонных условиях, размеры таких дефектов известны хорошо. Для других дефектов, взятых из образцов реальных трубопроводов, использовались прямые контрольные измерения и данные инспекции. Дополнительная информация о размерах также получалась от альтернативных дефектоскопических методов.

[2]Итого, нейросеть имеет 4 слоя - входы плюс три слоя нейронных процессорных элементов.

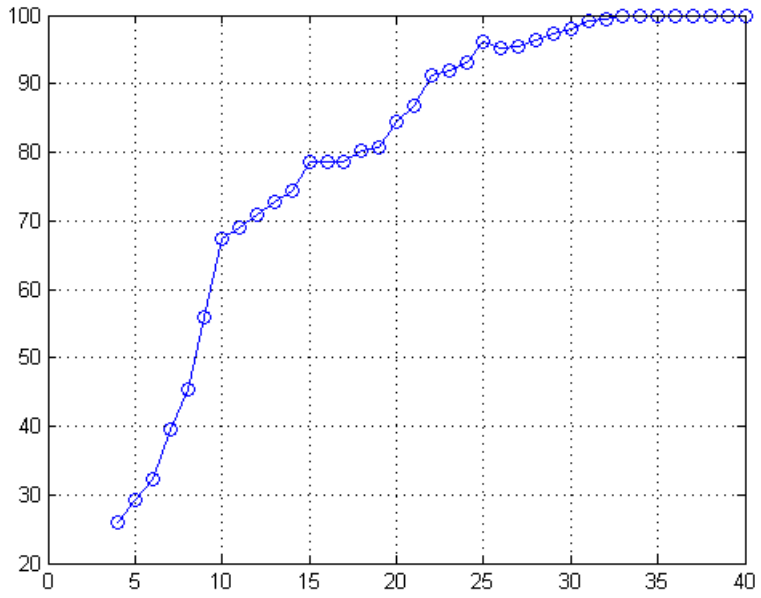

Рис. 2.2: Точность классификации дефектов по глубине (в %) в зависимости от объема выборки данных с метками. Использовано обучение нейронной сети с попутной аппроксимацией плотности данных.

регуляризация весов нейронов с априорным распределением Лапласа. Последовательный активный отбор обучающих меток основывался на энтропии выходов.

Для иллюстрации практических аспектов нейросетевого обучения была выбрана нейросеть избыточной для этой задачи сложности. Автоматическая регуляризация, как вследствие аппроксимации плотности, так и Байесовы[3] штрафы, приводит не только к отсутствию переобучения (на Рис. 6 достигается полное распознавание тестовых данных), так и к информативной структуре самой нейронной сети. Анализ весовых коэффициентов [4] выделяет информативные входы

[3]Автор считает, что понятия, связанные с такими именами, как Гаусс, Лаплас, Байес, вопреки складывающимся нормам языка, должны, по-крайней мере в учебно-научной литературе, начинаться с заглавной буквы.

[4]Анализ значений весов корректен и информативен, только если входы нормализованы и приведены к Гауссовому распределению применением выборочной обратной функцией для каждого входа.

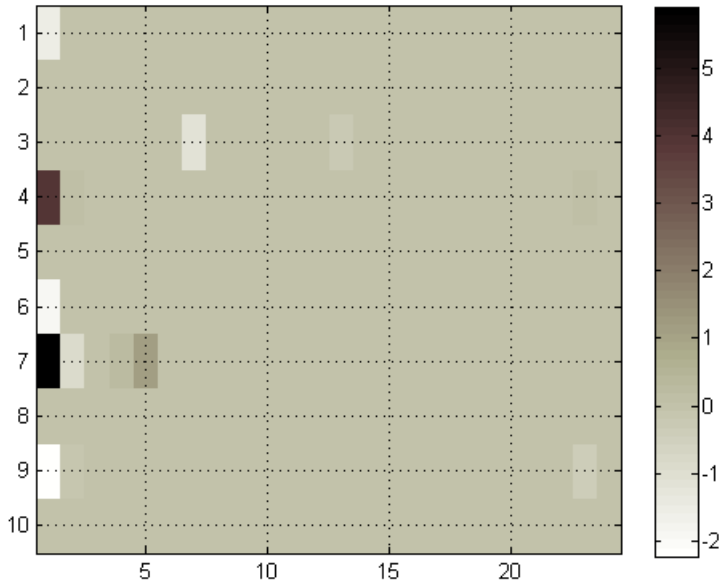

Рис. 2.3: Весовые коэффициенты входного слоя обученной нейронной сети с глубокой регуляризацией. Входы (по горизонтали) упорядочены по убыванию коэффициента линейной корреляции с выходами. Видно, что решение основано на использовании 7 входов из 24. При этом 4 из 10 нейронов практически не задействованы. Эта информация получается автоматически, вследствие корректной регуляризации.

и минимально достаточное число нейронов для классификации. Заметим, однако, что обучение нейронной сети, содержащей только это минимальное число нейронов, может оказаться гораздо более сложной вычислительной задачей оптимизации.

Основной результат - задача классификации успешно решена с использованием лишь 30-35 примеров с метками (около 5% от объема исходных данных). Эти примеры не произвольны, а выявлены путем направленного активного процесса анализа данных.

2.2 Итоги

В практических приложениях, как для промышленных потребителей, так и при интерпретации результатов научных экспериментов, полный корпус имеющихся данных, обычно, заметно богаче, чем выборка наблюдений с известными классами. Задача классификации фундаментальна, и для ее решения целесообразно использовать всю доступную информацию.

Затратный процесс сбора информативных меток классов необходимо оптимизировать. В лекции предложены подходы и методы оптимального отбора обучающих примеров, основанные на совмещении этого отбора с процессом последовательно уточняющегося обучения классификатора.

Направления активного обучения и гибридного обучения (с метками и без) бурно развиваются в последнее время. Эти исследования являются, по мнению автора, замечательным примером, когда теоретические построения и результаты напрямую сопрягаются с практическими потребностями пользователей обучающихся машин и алгоритмов.

Литература

[1] Т. Кохонен. Самоорганизующиеся карты. М. БИНОМ, 2010

[2] Xiaojin Zhu. Semi-Supervised Learning Literature Survey. Tech. Rep.TR 1530 University of Wisconsin – Madison, 2008.
URL: http://pages.cs.wisc.edu/∼jerryzhu/pub/ssl_ survey.pdf

[3] Olivier Chapelle, Alexander Zien, Bernhard Scholkopf (Eds). Semi-supervised learning. The MIT Press, 2006.

[4] Olivier Chapelle, Alexander Zien. Semi-Supervised Classification by Low Density Separation.
Max Planck Institute for Biological Cybernetics 72076 Tubingen, Germany.
URL: http://eprints.pascal-network.org/archive/00000388/01/pdf2899.pdf

[5] X.Li. Direct Solvers for Sparse Matrices.
URL: http://crd-legacy.lbl.gov/∼xiaoye/SuperLU/SparseDirectSurvey.pdf

[6] Geoffrey Hinton. Recent Developments in Deep Learning. Video lecture, 2010 (6-я минута).
URL: http://www.youtube.com/watch?v=VdIURAu1-aU

[7] G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. Science, 28 July 2006, Vol. 313, pp.504-507.

[8] С.Терехов. Нейросетевые аппроксимации плотности и их роль в информационном моделировании. Лекция для школы-семинара "Современные проблемы нейроинформатики Москва, МИФИ, 23-25 января 2002 года.
URL: http://alife.narod.ru/lectures/density2002/Density2002.pdf

[9] Warren B.Powell. Approximate Dynamic Programming. Wiley, 2011. (Глава 12)

[10] С.Терехов. История о многоруком бандите. Доклад Московскому клубу ”2015”, апрель 2011.
URL: http://neurokts.ru/docs/Multi_ Armed_ Bandit_ Story.pdf

[11] Interactive Machine Learning
URL: http://hunch.net/∼jl/projects/interactive/index.html

[12] A.J. Joshi, F. Porikli, N. Papanikolopoulos. Multi-class active learning for image classification. IEEE Conference on Computer Vision and Pattern Recognition, 2009. pp.2372-2379
URL: http://facweb.cs.depaul.edu/research/vc/seminar/Paper/final2[1].pdf

[13] С.Терехов. Гениальные комитеты умных машиню Лекция для школы-семинара "Современные проблемы нейроинформатики Москва, МИФИ, 24-26 января 2007 года.
URL: http://alife.narod.ru/lectures/committee2007/Committee2007.pdf

[14] Neil D. Lawrence , John C. Platt. Learning to Learn with the Informative Vector Machine. In Proceedings of the International Conference in Machine Learning, 2004.
URL: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.4533

[15] URL: http://en.wikipedia.org/wiki/Magnetic_ flux_ leakage

Глава 3

Приложения

3.1 Задачи

Об аппроксимации плотности

Задача восстановления плотности вероятности по конечной выборке данных является,очевидно, некорректно поставленной. В традиционных линейных методах оценивания (например, окна Парзена или ядерные оценки), решение сходится почти наверное (по вероятности), при росте объема выборки. Как следует понимать нелинейную нейросетевую аппроксимацию плотности? Следует ли сходимость таких оценок к плотности из теоремы об универсальных аппроксимирующих свойствах нейронных сетей?

О тестовых данных

При классификации с попутной аппроксимацией плотности, после построения классификатора широкое подмножество данных так и не получает меток классов. Таким образом, эти данные не задействованы напрямую для оценивания искомой условной вероятности $p(\text{выходы}|\text{входы})$. Однако, эти примеры без меток использованы при оценивании плотности для формирования базиса, на котором строится классификатор. Можно ли использовать эти примеры в качестве тестовой выборки, т.е., получив информацию об их метках, оценивать по этим примерам точность классификатора? Будет ли такая оценка смещенной (оптимистичной)?