УДК 004.93+519.2

Б.П. РУСЫН, В.А. ТАЯНОВ, А.А. ЛУЦЫК

ОЦЕНКИ НАДЕЖНОСТИ РАБОТЫ КЛАССИФИКАТОРОВ НА ОСНОВАНИИ ФУНКЦИИ НЕПОДОБИЯ

Ключевые слова: *верхняя оценка вероятности распознавания, скользящий контроль, устойчивость покрытия объектов классифицирующими алгоритмами, распределение расстояний, правая (левая) асимметрия, метрика, класс объектов.*

ВВЕДЕНИЕ

В настоящее время оценки вероятности правильного распознавания базируются на алгоритмах скользящего контроля (cross validation) [1]. Однако такие алгоритмы (исключение по одному и другие) трудоемкие с точки зрения вычислений и комбинаторных перегруппировок выборки. Поэтому необходимо разработать подходы к построению верхних оценок наибольшего значения средней вероятности правильного распознавания для значительно меньшего числа комбинаторных перегруппировок. Это возможно, поскольку обучающие данные практически всегда содержат избыток информации, что проявляется в ее частичном дублировании. С точки зрения переобучения, построение верхних оценок означает, что перед алгоритмом классификации поставлено наиболее сложную задачу (рассматривается выборка, на которой данный алгоритм будет ошибаться чаще, чем на остальных выборках из генеральной совокупности), включающую в себя произвольные более простые подвыборки [1, 2], т.е. оценки наибольшего значения средней вероятности правильного распознавания моделируют классификацию наиболее сложных подвыборок обучающей выборки. Чем меньше вероятность попадания в контрольную выборку более сложных подвыборок по сравнению с теми, которые входили в обучающую выборку, тем надежнее будет оценка. Вместе с тем при построении оценок для максимального значения средней вероятности правильного распознавания всей выборки, совокупности подвыборок, а также группы алгоритмов оценивается наибольшая вероятность устойчивого покрытия каждого объекта в отдельности. Под вероятностью устойчивого покрытия каждого объекта в отдельности понимается вероятность его правильной классификации при использовании в ряде подвыборок из генеральной совокупности объектов. В каждом из этих случаев вероятность вычисляется как средневзвешенное значение вероятностей по объектам с учетом вероятности их появления, а также их важности либо по алгоритмам (в алгоритмах голосования веса присваиваются алгоритмам в зависимости от надежности их экспертизы [3]), либо по обоим факторам. Таким образом, получается более точная полная оценка сверху для средней вероятности правильного распознавания.

В теории распознавания образов также принято оценивать сложность данных именно с точки зрения их классификации. В этом случае сложность данных оценивается на основании близости классов, к которым эти данные принадлежат, формы гиперповерхности их пересечения, а также форм их собственных поверхностей, количества данных классов, находящихся достаточно близко к разделяющей гиперповерхности, и тех, которые находятся по другую сторону разделяющей гиперповерхности, т.е. имеют отрицательный отступ [4].

ПОСТАНОВКА ЗАДАЧИ

В настоящей работе ставится общая проблема построения оценок вероятности правильной классификации объектов выборок, которая характеризует надежность классификации. Вначале рассматривается задача построения оценок вероятности правильной классификации каждого объекта в отдельности. Данная вероятность оценивается при условии, что известна принадлежность каждого рассматриваемого объекта к тому либо иному классу, т.е. рассматривается обучающая выборка. Эта вероятность может быть оценена на основании распределения отступов для соответствующих моделей либо распределением расстояний между объектами, в общем случае это одно и то же. Подробно данные вопросы рассматриваются ниже. Средняя вероятность правильной классификации всей выборки определяется как математическое ожидание от уже известных вероятностей правильной классификации каждого объекта в отдельности. При этом не обязательно знать априорные вероятности, которые зачастую неизвестны. Далее, основываясь на полукольце из операций $(\min, +)$ [5], можно показать, что среднее от верхних оценок правильной классификации каждого объекта в отдельности будет верхней оценкой средней вероятности правильной классификации всех объектов анализируемой выборки.

Затем рассматривается задача вычисления оценки правильной классификации в условиях скользящего контроля. В [1] анализируется полный скользящий контроль. Однако проводить его в условиях больших выборок очень трудоемкая задача с точки зрения вычислений. Полный скользящий контроль включает в себя многократную перегруппировку выборки методами скользящего контроля с возвратом по одному, с возвратом по k, оцениванием по блокам размерностью q, а также их перегруппировкой [4]. В связи с этим цель задачи — разработать метод построения верхних оценок для оценки полного скользящего контроля, используя скользящий контроль с небольшим числом перегруппировок выборки. В общем случае идея, заложенная в основу предлагаемого подхода, представляет собой оценку зависимости $f(x_2, x_3)$, если заданы зависимости $f(x_1, x_2)$ и $f(x_1, x_3)$. Как правило, решить эту задачу не представляется возможным, поэтому здесь описаны решения лишь для частных случаев. Поскольку в данной работе рассматриваются метрические классификаторы, предложен вариант решения задачи с помощью неравенства треугольника, что в общем случае представляет собой неравенство Коши–Шварца [6].

ВАЖНЫЕ ЗАДАЧИ ТЕОРИИ МАШИННОГО ОБУЧЕНИЯ

В современной теории машинного обучения существует две серьезные проблемы: получение точных верхних оценок вероятности нежелательного переобучения и способов его устранения. Под переобучением алгоритма распознавания подразумевается разность между вероятностями правильного распознавания при контроле и обучении. Оценивается вероятность того, что переобучение не превысит заданный вероятностный порог ε. На данный момент наиболее точные оценки сильно завышены. Экспериментально удалось установить причины этого. В порядке уменьшения влияния наиболее существенными из них являются следующие [1].

- **Пренебрежение эффектом расслоения или локализации семейства алгоритмов.** Данная проблема обуславливается тем, что реально работает не все множество алгоритмов, а только определенная его часть, зависящая от задачи. Коэффициент завышенности — от нескольких десятков до сотен тысяч.
- **Пренебрежение сходством алгоритмов.** Коэффициент завышенности — от нескольких сотен до десятков тысяч раз. Этот фактор всегда существенен и меньше зависит от задачи, чем первый.
- **Экспоненциальная аппроксимация «хвоста» гипергеометрического распределения.** Коэффициент завышенности может составлять несколько десятков.

• **Представление верхней оценки профиля разнообразия одним скалярным коэффициентом разнообразия.** Коэффициент завышенности часто порядка единицы, однако в некоторых случаях может достигать нескольких десятков.

Эффект переобучения состоит в том, что используется алгоритм с минимальным числом ошибок на обучающей выборке, т.е. проводится односторонняя настройка алгоритмов. Переобучение тем больше, чем большая композиция алгоритмов используется. Это справедливо для алгоритмов, взятых из распределения случайно и независимо. В случае зависимости алгоритмов (в реальной ситуации они, как правило, такими и являются) допускается уменьшение переобучения. Оно может возникнуть даже при выборе всего одного из двух алгоритмов. Расслоение алгоритмов по числу ошибок и увеличение их подобия уменьшают вероятность переобучения.

Рассмотрим дуплет выборка–алгоритм. Каждый алгоритм покрывает определенное число объектов обучающей выборки. Если использовать внутренние критерии [7] (например, в случае метрических классификаторов), то можно оценить устойчивость этого покрытия и сузить число покрытых объектов согласно заданному уровню устойчивости. Таким образом, для того чтобы покрыть большее число объектов, необходимо применить большее число алгоритмов. Эти алгоритмы должны быть похожими и иметь разный уровень ошибок. Это наилучшие современные стратегии построения композиции алгоритмов [1]. Однако при использовании тестовых данных, к которым композиция алгоритмов неадаптирована, ошибка классификации может сильно отличаться от минимальной, полученной на обучающих данных.

ПОСТРОЕНИЕ ОЦЕНОК ВЕРОЯТНОСТНОЙ УСТОЙЧИВОСТИ ПОКРЫТИЯ ОБЪЕКТОВ АЛГОРИТМАМИ ТИПА kNN ДЛЯ ОДИНОЧНЫХ ИСПЫТАНИЙ

Качество работы классификаторов, построенных на основании рангового голосования и с использованием разделяющих гиперплоскостей (R-моделей [3, с. 13]), принято характеризовать понятием отступа (margin), представляющем расстояние объекта от разделяющей гиперплоскости [4]. Чем больший отступ, тем лучшим считается классификатор. Понятие отступа применимо к классификаторам, построенным с помощью R-моделей, а также на основании функции подобия. К последним относятся все метрические классификаторы. Однако если все объекты или подавляющее их большинство имеют приблизительно одинаковый отступ и группируются один возле другого, то в этом случае резко падает их информативность. Это значит, что вместо всех объектов можно оставить один или несколько, используемых для обучения. Такой подход порождает одну из главных причин, обуславливающих переобучение. Односторонняя настройка алгоритма на основании близкой по сущности обучающей информации приводит к тому, что на контрольной выборке он может часто ошибаться, даже если не ошибался на обучающей выборке. Действительно, вероятность того, что в условиях обучающей выборки возможна такая же ситуация, близка к нулю.

Поэтому для обучения принято использовать непохожие и «трудные» для алгоритма объекты с малыми значениями отступа. Эта идея применяется, в частности, в методе опорных векторов (Support Vector Machine) или методе взвешенного голосования [3]. Используем обобщенный подход для характеристики классификаторов на основании понятия отступа. Результатом работы метрических классификаторов являются ранжированные данные (объекты базы данных, сортированные по степени подобия тестовому объекту). Для таких классификаторов понятие отступа представляется следующим образом. Вводится эквивалентная классическому отступу характеристика, которая может быть представлена как нормированное расстояние от тестового объекта до усредненного объекта

базы данных или последнего объекта из однородной (стратегической) [8] последовательности «своих» объектов. Допускается, что хотя бы часть «своих» объектов размещается в начале списка возможных претендентов. Таким образом, гарантируется корректность допущения.

Для более строгого определения данной характеристики необходимо ввести понятие распределения расстояний между объектами. Известно, что процесс с независимыми приращениями [9] имеет нормальное распределение с нулевым средним и линейно изменяющейся дисперсией. Каждое следующее приращение будет иметь большую дисперсию, чем предыдущее. Совокупность расстояний в признаковом пространстве от произвольного объекта до всех остальных является аналогом процесса с независимыми приращениями, если рассматривать попарные расстояния между объектами выборки, дисперсия которого ограничивается замкнутостью признакового пространства. Распределение расстояний существует, поскольку существует распределение независимых приращений [9]. Поскольку произвольное расстояние — это всегда положительная величина, то ее математическое ожидание больше нуля, если оно не равно нулю.

Чтобы перейти к нормальному распределению с нулевым математическим ожиданием, необходимо сделать сдвиг влево всего распределения на величину, равную математическому ожиданию. Закон распределения совокупности расстояний определяется согласно центральной предельной теореме, из которой следует, что сумма произвольных случайных величин будет иметь асимптотически нормальное распределение [10]. Математическое ожидание и дисперсию плотности распределения вероятностей можно оценить, например, методом максимального правдоподобия [11].

Поскольку значения расстояний могут быть произвольными по модулю, то процедура непараметрического оценивания формы закона распределения неусеченными ядерными функциями будет корректной.

Пусть непараметрически оценена плотность распределения расстояний между объектами, заданными векторами $\mathbf{x}$ и $\mathbf{y}$: $p(x)$, $x \to d(\mathbf{x}, \mathbf{y})$. Согласно неравенству Чебышева [12] вероятность того, что найдется расстояние, превышающее некоторое пороговое значение расстояний θ, равна $\int\limits_{|x|\ge\theta} p(x)dx \le \dfrac{\sigma^2}{\theta^2}$.

Рассмотрим случай равенства математического ожидания и моды распределения $p(x)$. Верхний предел одномодального распределения с модой $\mu = 0$ неравенством Гаусса [13] представляется в виде

$$P(|x-\mu| \ge \lambda\tau) \le \frac{4}{9\lambda^2}, \tag{1}$$

где $\tau^2 \equiv \sigma^2 + (\mu - \mu_0)^2$.

Пусть $\mu = \mu_0 = 0$ и $\tau \equiv \sigma$. Тогда порог $\theta = \lambda\tau = \lambda\sigma$, а $\lambda = \theta/\sigma$. Итак, неравенство Гаусса для порога θ может быть представлено в виде

$$\int\limits_{|x|\ge\theta} p(x)dx \le \frac{4\sigma^2}{9\theta^2}. \tag{2}$$

Таким образом, согласно неравенству Гаусса для одномодальных распределений с модой, равной математическому ожиданию, оценка в 2,25 раза лучше той, которая получается согласно неравенству Чебышева. Это максимально хорошая оценка при условии, что неизвестен конкретный вид распределения, а известны лишь определенные его свойства. При симметричности одномодального распределения мода равна математическому ожиданию, а в частном случае нормированности оба параметра равны нулю. Однако в общем случае иногда реальный закон распределения не является симметричным. При этом возможна левая либо правая асимметрия функции плотности распределения вероятностей (ФПРВ).

Разделим ФПРВ на две части: находящиеся справа и слева от максимума. Если площадь под правой частью ФПРВ больше левой, то считается, что это правая асимметрия, а если наоборот, то левая (рис. 1, *а* и *б* соответственно). Рассмотрим оценки, полученные с помощью неравенства Гаусса для обоих случаев. Для правой асимметрии сделаем распределение симметричным относительно левой части, т.е. левую часть оставляем без изменений и отображаем ее симметрично вместо исходной правой части. Пусть некоторая точка x_0 принадлежит левой части распределения. Тогда функция распределения вероятностей (ФРВ) $P(X < x_0)$ для симметричного случая всегда больше для каждой точки в левой части ФРВ, нежели в исходном случае. Нас интересуют первые объекты в списке возможных претендентов, соответствующие левой части распределения. Именно это дает основание для искусственной симметризации закона распределения, являющегося в общем случае асимметричным. При этом ФРВ будет верхней оценкой для ошибки распознавания.

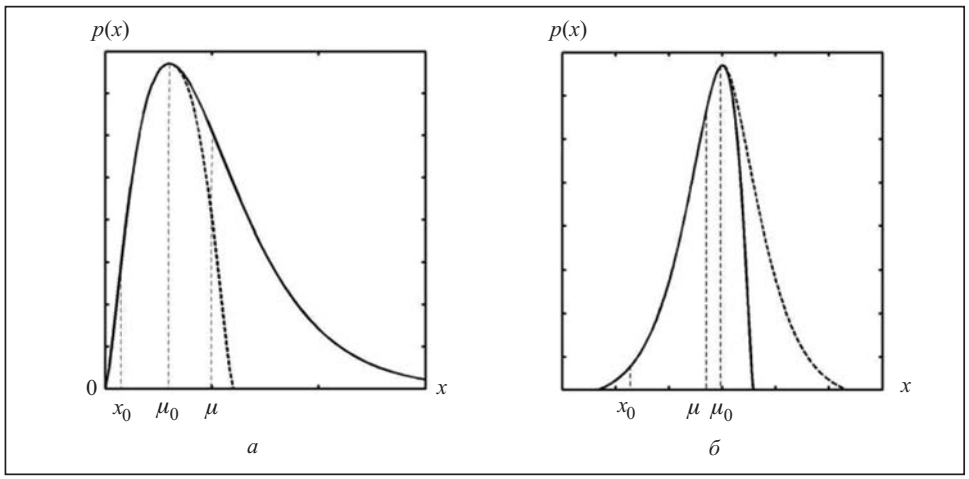

Рис. 1

Проанализируем полученный результат. Предварительно отметим, что для лучшего понимания предложенного приема, а также облегченной интерпретации результатов нет необходимости в нормировке ФПРВ к единичной площади. Отметим, что оценка дисперсии может быть проведена только по одной части распределения (в данном случае — левой). Это связано с тем, что объекты части распределения, которые не участвуют в определении оценки дисперсии, значительно удалены от зоны принятия решений (в данном случае — первых объектов в списке возможных претендентов) и не оказывают существенного влияния на принимаемое решение. Уменьшение единичной площади свидетельствует о том, что мы получаем более точную оценку, нежели по всему распределению. Увеличение площади, наоборот, свидетельствует об ухудшении оценки.

Особый интерес представляют отклонения расстояний влево от математического ожидания при использовании kNN классификаторов с небольшими значениями k. Поскольку оценка дисперсии ФПРВ для построения оценки Гаусса проводилась по левой части распределения, то очевидно, что эта оценка в случае симметричной ФПРВ меньше исходной, что делает оценку более точной. К тому же симметрия позволяет сделать оценку Гаусса максимально точной согласно неравенству (2), а все вместе позволяет существенно улучшить общую верхнюю оценку.

Рассмотрим ФПРВ в случае левой асимметрии. Теперь дисперсия симметризированной ФПРВ будет больше исходной, а единственным преимуществом такого преобразования будет симметрия вновь полученного закона распределения. В данном случае также нет необходимости в нормировке ФПРВ. Увеличение площади под кривой означает, что включены дополнительные объекты, которые не участвуют в распознавании. Это ухудшает оценку Гаусса, поскольку возросло значение оцененной дисперсии. Решение о том, какую оценку использовать — с преобразованием симметрии или по исходному распределению, необходимо принимать, имея значения математического ожидания, моды и дисперсии обеих ФПРВ.

Проанализируем связь оценки Гаусса со значениями ФРВ $P(X < x_0)$. Правая часть распределения не представляет интереса, поэтому если вместо оценки Гаусса взять ФРВ, то это будет верхней оценкой по отношению к исходной оценке. При этом не имеют значения ни вид асимметрии, ни сама асимметрия в ФПРВ. Итак, верхняя оценка значениями ФРВ по отношению к оценке Гаусса касается как симметричных, так и асимметричных ФПРВ. Завышенность оценки Гаусса по отношению к значениям ФРВ, безусловно, компенсируется лишь в случае правой асимметрии. В случае левой асимметрии степень компенсации зависит от соотношения между значениями дисперсии и разницы $|\mu - \mu_0|$.

Если ФПРВ не имеет четко выраженной структуры (существование экстремума, симметрия, правая асимметрия), то можно воспользоваться непараметрическим оцениванием, в результате которого получаем непрерывную ФПРВ. Эту функцию можно интегрировать и дифференцировать по определению. Поскольку нормальная ФПРВ характеризуется минимальной ошибкой классификации для данного порога θ и не превышает $\dfrac{4\sigma^2}{9\theta^2}$ [13, 14] в случае одномодальной симметричной ФПРВ либо ФПРВ с правой асимметрией, двустороннее неравенство для данной ошибки распознавания запишем

$$0{,}5\left(1 - erf\left(\frac{\theta}{\sigma}\right)\right) \le \varepsilon \le \frac{4\sigma^2}{9\theta^2}, \tag{3}$$

где $\mu = 0$.

Проанализируем общую возможную форму потенциально получаемых ФПРВ расстояний между объектами. Все распределения будут иметь экстремумы, поскольку ФПРВ существует на интервале $[0, \infty)$, а плотность в окрестности нуля и для больших расстояний не может быть высокой, так как эти события маловероятны. Правая асимметрия (см. рис. 1, *а*) более вероятная, поскольку ФПРВ расстояний ограничена нулем и не имеет строгих ограничений.

ОЦЕНКИ ВЕРОЯТНОСТНОЙ УСТОЙЧИВОСТИ ПОКРЫТИЯ ОБЪЕКТОВ АЛГОРИТМАМИ ТИПА *k*NN В УСЛОВИЯХ ДВУХ КЛАССОВ, ИМЕЮЩИХ ЗАДАННЫЕ РАЗМЕРЫ

Рассмотрим распространенную задачу классификации в условиях двух классов. Обозначим размеры классов s_1 и s_2. Тогда если вероятность замещения объекта из класса размером s_1 в пределах доверительного интервала равна ε_1, то вероятность незамещения объектов из этого же класса объектами из класса размером s_2 равна $(1-\varepsilon_1)^{s_2}$ при условии независимости объектов [15]. Для другого класса при соответствующих изменениях в обозначениях эта вероятность равна $(1-\varepsilon_2)^{s_1}$. Если ввести некоторый виртуальный класс и допустить, что замещение произвольного объекта этого класса объектами из упомянутых двух классов является достоверным событием, то можно записать следующее уравнение:

$$\gamma((1-\varepsilon_1)^{s_2} + (1-\varepsilon_2)^{s_1}) = 1, \tag{4}$$

откуда множитель пропорциональности γ вычисляется тривиально.

Иногда имеют место ситуации, когда расстояния между объектами равны нулю. При этом непараметрически оцененное распределение одного из классов может иметь максимум в точке, соответствующей нулевому расстоянию. Пусть плотности распределений в нулевой точке равны $p_1(0)$ и $p_2(0)$. Оценка соотношения между вероятностями может быть задана в виде $p_1(0)^{s_2} / p_2(0)^{s_1}$ или $\ln (p_1(0)^{s_2} / p_2(0)^{s_1})$. При этом необходимо сделать граничный переход от ФРВ к ФПРВ, поскольку они связаны между собой операцией дифференцирования. Соотношение $\ln (p_1(0)^{s_2} / p_2(0)^{s_1})$ ($\ln (p_2(0)^{s_1} / p_1(0)^{s_2})$ или в общем случае $\ln (p_1(\theta)^{s_2} / p_2(\theta)^{s_1})$ ($\ln (p_2(\theta)^{s_1} / p_1(\theta)^{s_2})$) можно использовать для построения классификатора вида

$$\ln \frac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}} > \gamma_1;\ \ln \frac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}} < \gamma_1 \ \text{ или } \ \ln \frac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}} > \gamma_2;\ \ln \frac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}} < \gamma_2, \quad (5)$$

где значение $\ln \dfrac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}} = 0$ или $\ln \dfrac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}} = 0$ не влияет на результаты классификации, а решение может быть принято в пользу любого класса. В случае непараметрического оценивания вероятность такого значения практически равна нулю.

ВЕРХНЯЯ ОЦЕНКА ДЛЯ МЕТОДА СКОЛЬЗЯЩЕГО КОНТРОЛЯ МЕТРИЧЕСКИХ АЛГОРИТМОВ КЛАССИФИКАЦИИ

Методы скользящего контроля неразрывно связаны с таким понятим, как обобщающая способность алгоритмов классификации. Под обобщающей способностью алгоритмов подразумевается их способность относить объекты со схожими свойствами к одному и тому же классу. Качество алгоритмов, с точки зрения обобщающей способности, принято характеризовать на основании частоты (вероятности) ошибок, которые совершает тот или иной алгоритм. При этом важно также предвидеть частоту ошибок на контрольной выборке, к которой алгоритм классификации неадаптирован. Частоту ошибок на контрольной выборке можно предвидеть на основании гипотезы о независимости объектов выборки (которая обычно всегда выполняется [1]), а также обобщающей способности алгоритмов.

Если существует тесная связь между частотой ошибок при обучении и на контроле, то, уменьшая частоту ошибок при обучении (часто она практически равна нулю), мы сможем уменьшить ее также и на контроле или предвидеть ее с большой вероятностью.

Пусть X — пространство объектов (object space); Y — множество имен классов (class name set); $y^*: X \to Y$ — целевая функция (target function), значения которой известны лишь на объектах конечной обучающей выборки длины l: $X^l = (x_i, y_i)_{i=1}^l \subset X \times Y$, $y_i = y^*(x_i)$. В базе данных существуют классы эталонов (class patterns) C_i, $i = \overline{1, n}$, причем $s_i = |C_i|$ — размеры классов. Предполагается, что размеры s_i всех классов одинаковые и равны s. Поскольку существует выборка контрольных объектов U, подающихся на распознавание, то общее количество объектов, участвующих в процессе распознавания, равно $n \cdot s + |U|$. Пусть оцененная частота ошибок (error frequency) алгоритма классификации $a = \mu(X^l)$ на обучающей выборке $X^l \subseteq X^L$: $\nu(a, U) = \dfrac{1}{|U|} \sum_{x \in U} [a(u) \neq y^*(u)]$, где запись $x \in U$ означает, что объект относится к контрольной последовательности, а запись $[a(u) \neq y^*(u)]$ воспринимается как функция индикации несовпадения ответа алгоритма $a(u)$ и правильного ответа $y^*(u)$ для этого объекта.

На практике оптимальное значение k подбирается по критерию скользящего контроля (cross-validation) с исключением объектов по одному (leave-one-out, LOO). Для каждого объекта $x_i \in X^l$ проверяется, правильно ли он классифицируется по своим k ближайшим соседям:

$$\mathrm{LOO}(k, X^l) = \sum_{i=1}^{l} [a(x_i; X^l \setminus x_i, k) \neq y_i] \to \min_k. \quad (6)$$

Идея предлагаемого подхода состоит в следующем. Анализируются расстояния между тестовым объектом и объектами базы данных. Тогда на основании расстояний от тестового объекта до объектов базы данных необходимо предвидеть расстояния между объектами базы данных. На первый взгляд может показаться, что такие оценки менее точны, нежели получаемые на основании полного скользящего контроля по выборке. Однако если доказать, что эти оценки представляют собой верхние оценки относительно частоты (вероятности) ошибок, то они будут менее чувствительны к факторам, обуславливающим переобучение.

Пусть степень подобия между объектами характеризуется на основании понятия расстояния. По определению расстояние между двумя векторами признаков (x и y) соответствующих объектов должно удовлетворять следующим условиям:

1) $d(x, x) = d(y, y) = 0$;

2) $d(x, y) = d(y, x)$

(для того чтобы расстояние было метрикой, необходимо выполнение еще одного дополнительного условия, называемого неравенством треугольника):

3) $d(x, y) \le d(x, z) + d(y, z)$.

Метрикой является обобщенная метрика Минковского с показателем степени $p \ge 1$:

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} = \left(\sum_{i=1}^{n} a_i |x_i - y_i| \right)^{1/p} = C(p) \sum_{i=1}^{n} a_i |x_i - y_i|, \quad (7)$$

где мультипликативный множитель $C(p)$ представляется в виде

$$C(p) = \left(\sum_{i=1}^{n} a_i |x_i - y_i| \right)^{(1-p)/p}, \; a_i = |x_i - y_i|^{p-1}, \; p > 0. \quad (8)$$

Проанализируем расстояния между объектами с точки зрения адекватного описания подобия между объектами, а также метрики Минковского для различных показателей p. Введем понятие глубины метрики $M = \dfrac{d(x, z) + d(y, z)}{d(x, y)}$. Таким образом, чем больше значение глубины метрики, тем строже выполняется условие треугольника. Для метрики Минковского ее глубина возрастает с ростом показателя p. По определению $M \ge 1$. Остается выяснить, как влияет та или иная метрика на перераспределение расстояний между объектами.

Сначала рассмотрим евклидовую метрику (показатель в обобщенной метрике Минковского равен 2): $d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2}$. Пусть известны расстояния между тестовым объектом и объектами базы данных: $d_i = d(x, y_i)$ и $d_j = d(x, y_j)$. Необходимо оценить расстояние $d_{ij}(y_i, y_j)$ на основании двух известных расстояний. В теории распознавания образов или теории машинного обучения эта задача сводится к оценке результатов на контроле, когда известны результаты лишь на обучающей выборке. Количество данных, используемых для обучения, значительно меньше того количества данных, которое можно было бы получить

комбинаторной перегруппировкой выборки. Это нужно для быстрого определения параметров без учета результатов распознавания алгоритмами полного скользящего контроля, т.е. проводится неполный скользящий контроль с глубиной, значительно меньше той, что может быть получена перегруппировкой общей выборки на тестовую и обучающую.

Рассмотрим в качестве примера три расстояния: $d_1 = d(x, y_1)$, $d_2 = d(x, y_2)$ и $d_3 = d(x, y_3)$ для евклидовой метрики и пространства R^2. Соответственно необходимо оценить расстояния $d_{12} = d(y_1, y_2)$, $d_{13} = d(y_1, y_3)$ и $d_{23} = d(y_2, y_3)$. Вычислим их с помощью теоремы косинусов:

$$\begin{aligned} d_{12} &= d_1^2 + d_2^2 - 2d_1 d_2 \cos(d_1, d_2), \\ d_{13} &= d_1^2 + d_3^2 - 2d_1 d_3 \cos(d_1, d_3), \\ d_{23} &= d_2^2 + d_3^2 - 2d_2 d_3 \cos(d_2, d_3). \end{aligned} \tag{9}$$

Для того чтобы определить расстояния d_{12}, d_{13} и d_{23}, необходимо знать соответствующие углы между векторами. Возможны два варианта:

1) применить параллельно евклидовой косинусную метрику;

2) вычислить завышенную оценку для соответствующих расстояний, накладывая определенные условия на углы между векторами.

Первый способ гарантирует точное вычисление расстояний лишь в случае евклидовой метрики. Второй способ используется лишь как базовый для евклидовой метрики, что в дальнейшем позволит оценивать соотношения между расстояниями с помощью других метрик. Часто нет необходимости определять расстояние абсолютно точно, чтобы результаты распознавания были идентичными; кроме того, такая оценка дает запас устойчивости результатов распознавания. Поэтому рассмотрим второй вариант. Допустим, что углы между векторами d_1, d_2 и d_3 находятся в пределах $[0; \pi/2]$. Это вытекает из того, что если угол равен $\pi/2$, достигается максимальное различие между объектами, описываемыми соответствующими векторами. При этом для того чтобы больший из возможных углов между векторами расстояний был равен $\pi/2$, хотя бы один из углов, образованных векторами расстояний d_1, d_2 и d_3 с положительной полуосью, должен быть не меньше $\pi/2$. Пусть для классификации заданной последовательности объектов важными являются k ближайших соседей. Проанализируем процесс принятия решения на основании алгоритма ближайших соседей. Для него расстояния d_{12}, d_{13} и d_{23} могут быть вычислены следующим образом:

$$\begin{aligned} d_{12} &= d_1^2 + d_2^2 - 2d_1 d_2 \cos\left(\frac{\pi}{2l}\right), \\ d_{13} &= d_1^2 + d_3^2 - 2d_1 d_3 \cos\left(\frac{\pi}{2l}\right), \\ d_{23} &= d_2^2 + d_3^2 - 2d_2 d_3 \cos\left(\frac{\pi}{2l}\right). \end{aligned} \tag{10}$$

Этот способ вычисления расстояний основан на допущении о равномерности распределения углов между соответствующими векторами расстояний. Для задачи распознавания такое допущение соответствует наихудшему случаю. Итак, на основании (10) получены верхние оценки для соответствующих расстояний. Теперь эти расстояния для произвольной последовательности объектов и векторов, которые им соответствуют, запишем так:

$$d_{ij} = d_i^2 + d_j^2 - 2d_i d_j \cos\left(\frac{\pi(j-i)}{2l}\right), \quad i, j \in \{1, k\}; \quad j > i. \tag{11}$$

Рассмотрим изменение расстояний между объектами при применении других метрик, учитывая два возможных случая по отношению к евклидовой метрике. Первый касается метрик с порядком $p > 2$, а второй — $1 \le p < 2$. Проанализируем изменение соотношения между векторами расстояний на основании понятия глубины метрики M. С ростом показателя p в обобщенной метрике Минковского глубина метрики M растет. При этом для случая $p > 2$ глубина метрики всегда больше, чем для $p = 2$, а для $1 \le p < 2$ — соответственно меньше. Таким образом, евклидовая метрика является некоторой границей раздела в пространстве метрик, относительно которой проводится сравнение. Поэтому евклидовая метрика наиболее используемая в прикладных задачах, а расстояния, вычисляемые с ее помощью, понятны с точки зрения интерпретации.

ЗАКЛЮЧЕНИЕ

В настоящей работе построены и исследованы оценки вероятности правильной классификации для классификаторов, использующих в качестве меры подобия функцию расстояний. Результаты оценивания получены на основании функции распределения расстояний между объектами. При этом рассмотрены разные частичные случаи формы функции распределения. Построены двусторонние верхние оценки одиночного распознавания и распознавания для двух классов заданных размеров. Предложен метод классификации на основании соотношения плотностей распределения вероятностей в нулевой и произвольных точках. Разработан подход к построению оценок максимального значения вероятности правильного распознавания для классификаторов на основании функции расстояний (классификаторов типа kNN) с помощью неглубокого скользящего контроля.

СПИСОК ЛИТЕРАТУРЫ

1. Vorontsov K.V. Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. — 2008. — **18**, N 2. — P.243–259.
2. Vapnik V. The nature of statistical learning theory. — New York: Springer-Verlag, 2000. — 314 p.
3. Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. — 1978. — **33**. — С. 5–68.
4. Воронцов К.В. Машинное обучение и анализ данных // Курс лекций «Математические методы обучения по прецедентам». — http: // www.ccas.ru/voron/teaching.html.
5. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наук. думка, 2004. — 545 с.
6. Moon T.K., Stirling W.C. Mathematical methods and algorithms for signal processing. — N.J.: Prentice-Hall, 2000. — 937 p.
7. Kapustii B.E., Rusyn B.P., Tayanov V.A. Classifier optimization in small sample size condition // Automatic Control and Computer Sci. — 2006. — **40**, N 5. — P. 17–22.
8. Капустій Б.О., Русин Б.П., Таянов В.А. Комбінаторна оцінка впливу зменшення інформаційного покриття класів на узагальнюючу властивість 1NN алгоритмів класифікації // Искусственный интеллект. — 2008. — № 1. — С. 49–54.
9. Карлин С. Основы теории случайных процессов. — М.: Мир, 1971. — 576 с.
10. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. — М.: Наука, 1985. — 640 с.
11. Bishop C.M. Pattern recognition and machine learning (Information science and statistics). — London: Springer, 2006. — 738 p.
12. Weisstein E.W. Chebyshev inequality. — http://mathworld.wolfram.com/ChebyshevInequality.html, 10.12.2008.
13. Weisstein E.W. Gauss inequality. — http://mathworld.wolfram.com/GaussInequality.html, 10.12.2008.
14. Ту Дж., Гонсалес Р. Принципы распознавания образов. — М.: Мир, 1978. — 416 с.
15. Капустий Б.О., Русын Б.П., Таянов В.А. Новый подход к определению вероятности правильного распознавания объектов множеств // УСиМ. — 2005. — № 2. — С. 8–13.

Поступила 07.12.2011