УДК: 681.518.2

М.Н. Нессонова

МЕТОД РЕЙТИНГОВОГО ГОЛОСОВАНИЯ КОМИТЕТА АЛГОРИТМОВ В ЗАДАЧАХ КЛАССИФИКАЦИИ С УЧИТЕЛЕМ

Национальный фармацевтический университет, г. Харьков

Ключевые слова: *классификация с учителем, дискриминация, решающие правила, классификатор, комитет (ансамбль, смесь, композиция) алгоритмов.*

Предложен алгоритм метода голосования комитета классификаторов в задаче построения решающих правил дискриминации объектов на несколько классов, основанный на вычислении рейтингов принадлежности классифицируемого образца к каждому классу, расчет которых учитывает как точность алгоритмов смеси на отдельных классах, так и их ошибки на других классах.

М.М. Нессонова

Метод рейтингового голосування комітету алгоритмів у задачах класифікації з учителем

Ключові слова: *класифікація з учителем, дискримінація, вирішальні правила, класифікатор, комітет (ансамбль, суміш, композиція) алгоритмів.*

Запропоновано алгоритм методу голосування комітету класифікаторів у завданні побудови вирішальних правил дискримінації об'єктів на кілька класів, що ґрунтується на обчисленні рейтингів належності зразка, що класифікується, до кожного класу, розрахунок яких враховує як точність алгоритмів суміші на окремих класах, так і їх помилки на інших класах.

M.M. Nessonova

Method of rating voting of algorithms committee in classification tasks with teacher

Key words: *classification with a teacher, discrimination, decision rules, classifier, committee (ensemble, mixture, composition) of algorithms.*

The algorithm of method of voting of committee of classifiers is offered in the task of construction of decision rules of discrimination of objects into some classes, based on the calculation of rating of belonging of the classified standard to every class, the calculation of that takes into account both exactness of algorithms of mixture on separate classes and their errors on other classes.

Задачи классификации с учителем (дискриминации) в медицинских приложениях возникают при дифференциальной диагностике заболеваний, при оценке степени и тяжести состояния пациентов, при анализе генных структур, а также в других актуальных направлениях теоретических и практических исследований. Основная цель при решении этого типа задач состоит в построении правил (называемых классификаторами, или алгоритмами классификации) отнесения объектов к одной или нескольким заранее известным группам. Решаются задачи как одноклассовой классификации объектов (например, когда по некоторому набору признаков необходимо отличить конкретное заболевание от всех других возможных состояний пациента), так и классификации наблюдений на несколько классов (например, когда необходимо оценить форму заболевания как легкую или тяжелую).

Постановка задачи классификации с учителем в общем виде выглядит следующим образом. Пусть есть множество объектов X и конечный набор классов $\{C_i\}_{i=1}^{n}$. Известно, что каждый объект $x \in X$ относится к некоторому классу $C_j \in \{C_i\}_{i=1}^{n}$. Необходимо построить правило (алгоритм) $T: X \to \{C_i\}_{i=1}^{n}$, относящее объекты из X к некоторому (своему) классу. Обычно в реальности исследователь не имеет в своем распоряжении сведений ни обо всех объектах, которые необходимо классифицировать, ни исчерпывающих характеристик этих объектов. Поэтому для объектов используют некоторый конечный набор их признаков (количественных или/и качественных), а подгонка параметров алгоритма (его обучение) проводится по некоторому конечному подмножеству $X^L \subset X$, называемому обучающей выборкой. Проблема построения классификатора наилучшей точности, сохраняющего при этом хорошую обобщающую способность, является здесь основной, и для ее решения используются различные методы обучения алгоритмов. Однако при решении задач часто возникает ситуация, когда ни увеличение числа признаков, описывающих объект, ни выбор алгоритма из различных семейств, ни применение к нему более эффективных методов обучения в итоге не дают классификатора с достаточно приемлемой точностью. Например, одно решающее правило дает хорошие результаты при дискриминации объектов первого класса и неудовлетворительную точность определения принадлежности к другим классам; другие классификаторы – с большой точностью отделяют объекты других классов, имея низкую точность на объектах первого класса. Выходом

здесь может быть объединение нескольких алгоритмов в композицию (комитет, ансамбль) с целью компенсации их взаимных ошибок.

Используют следующие способы построения композиций алгоритмов:

1) Голосование по большинству (простое голосование), при котором комитет классификаторов относит объект к тому классу, к которому его отнесли большинство входящих в него алгоритмов;

2) Голосование по старшинству (машина покрывающих множеств). Этот метод предполагает последовательную одноклассовую классификацию. Т.е. первый алгоритм комитета отвечает за отнесение объекта к классу 1. Если он отказывается от классификации, то объект передается второму алгоритму, который может его отнести к классу 2. Если этого не произошло, объект передается к третьему классификатору и т. д., пока один из алгоритмов не примет решения.

3) Взвешенное голосование и смеси экспертов. Здесь голос каждого из классификаторов $\{T^k\}^m_{k=1}$, входящих в комитет T, имеет свой вес α_k, зависящий от ошибки данного алгоритма на обучающем множестве:

$$T(x) = \sum_{k=1}^{m} \mathsf{a}_k \cdot T^k(x)$$

В случае $\alpha_k=\ln(1-/pk)$, $k=1,...m$, где pk – ошибка классификатора T^k, получаем простейший линейный «наивный» байесовский классификатор. В случае $\alpha_1=\alpha_2=...=\alpha_m=1/k$ имеем дело с голосованием по большинству. Если же весовые коэффициенты $\alpha_k=\alpha_k$, $k=1,...m$, т. е. являются не постоянными, а зависят от самого́ классифицируемого объекта, то комитет T называют смесью экспертов (алгоритмов), а функции $\alpha_k(x)$ – шлюзами, или функциями компетентности.

ЦЕЛЬ РАБОТЫ

Построение алгоритма процедуры голосования смеси экспертов в задаче классификации на несколько классов, основанный на вычислении рейтингов принадлежности классифицируемого образца к тому или иному классу.

При расчете рейтингов наравне с точностью алгоритмов смеси при прогнозировании отдельных классов учитываются и ошибки этих классификаторов на других классах.

Имеется конечное множество из m классификаторов $\{T^k\}^m_{k=1}$, решающих задачу дискриминации объектов $x \in X$ на n классов $\{C_i\}^n_{i=1}$. Для всех T^k известна (например, оценена по обучающей выборке) их точность прогнозирования классов, которая может быть задана квадратными $n\times n$-матрицами $P^k=(p^k_{ij})^n_{i,j=1}$ где каждый элемент p^k_{ij} рассматривается как вероятность того, что объект, классифицированный алгоритмом T^k, как принадлежащий к классу C_j, в действительности принадлежит к классу C_i. Таким образом, диагональные элементы p^k_{ij} характеризуют точность алгоритмов T^k на классах C_i $(\forall k = 1,...m; \quad \forall i = 1,...n)$,

а сумма элементов по каждому столбцу матриц равна единице:

$$\sum_{i=1}^{n} p^k_{\mathrm{ij}} = 1, \quad \forall j = 1,...n, \quad \forall k = 1,...m$$

Результат действия множества построенных алгоритмов $\{T^k\}^m_{k=1}$ на некотором объекте x представляется бинарной индикаторной $n\times m$-матрицей

$$V(x) = \left(v_{\mathrm{ij}}(x)\right)^{j=1,...m}_{i=1,...n}, \text{ где } v_{\mathrm{ij}}(x) = \begin{cases} 1, & \text{если} \quad T^j(x) = C_i \\ 0, & \text{если} \quad T^j(x) \neq C_i \end{cases}$$

Если каждый классификатор дает на новом объекте один и только один ответ, то

$$\sum_{i=1}^{n} v_{\mathrm{ij}}(x) = 1, \quad \forall j = 1,...m.$$

это означает, что в каждом столбце матрицы V имеется только один ненулевой элемент. При условии, что некоторые алгоритмы могут отказываться от классификации некоторых объектов, в матрице V возможно появление столбцов, целиком состоящих из нулей. Т.е. в общем виде

$$\sum_{i=1}^{n} v_{\mathrm{ij}}(x) \le 1, \quad \forall j = 1,...m, \quad \forall x \in X.$$

Далее формируем т. н. «рейтинговую» $n\times m$-матрицу W, каждый столбец которой получается умножением матрицы P^k на k-й столбец матрицы V:

$$W^{\langle k \rangle}(x) = P^k \cdot V^{\langle k \rangle}(x).$$

Таким образом, элементы матрицы W $\forall i = 1,...n, \ \forall k = 1,...m$ определяются соотношением:

$$w_{\boldsymbol{k}}(x) = \begin{cases} p^k_{i\mathsf{q}}, & \text{если} \quad T^k(x) = C_{\mathsf{q}}, \quad \text{где} \quad \mathsf{q} \in \{1,...n\} \\ 0, & \text{если} \quad T^k(x) \quad \text{отказывается от классификации} \end{cases}.$$

Рейтинг принадлежности объекта x к классу C_i вычисляется как сумма по i-й строке матрицы W:

$$r_{C_i}(x) = \sum_{k=1}^{m} w_{\mathrm{ik}}(x) = \sum_{k=1}^{m} p^k_{i\mathsf{q}}, \text{ где } \mathsf{q} \in \{0,1,...n\}$$

– номер класса, к которому относит объект x алгоритм T^k (при этом под $\theta = 0$ подразумевается отказ алгоритма от классификации данного объекта).

Итоговым результатом процедуры рейтингового голосования комитета классификаторов T является класс, которому соответствует наибольший рейтинг: $T(x)=C_t$,
где $\mathsf{t} = \arg\max_{i=1,...n} r_{C_i}(x)$.

ВЫВОДЫ

Предложенная процедура рейтингового голосования смеси экспертов позволит повысить надежность ответа, выдаваемого ансамблем классификаторов в целом.

Сведения об авторе:
Нессонова М.Н., ассистент каф. фармакоинформатики НФаУ.

Поступила в редакцию 15.10.2012 г.