Обучение композиций дипольных классификаторов на основе ЕМ-алгоритма

Пустовойтов Н. Ю.
pustovoytov@forecsys.ru
Москва, МФТИ, ЗАО «Форексис»

Идея дипольного классификатора, предложенная в [2], так же, как и идея метода потенциальных функций [1], заимствована из физики. Диполем называется пара зарядов, имеющих одинаковый по модулю, но противоположный по знаку заряд, и расположенных на незначительном расстоянии друг от друга. Диполь разделяет пространство плоскостью на области положительного и отрицательного потенциала. Поэтому его можно рассматривать как линейный классификатор, заданный двумя точками (полюсами диполя), либо как алгоритм ближайшего соседа, построенный по выборке из двух объектов (полюсов). В данной работе рассматриваются композиции дипольных классификаторов.

Дипольный классификатор

Пусть $X \subset \mathbb{R}^n$ — множество *объектов*, $Y = \{-1, +1\}$ — множество *меток классов*. Требуется построить алгоритм классификации $a\colon X \to Y$, восстанавливающий неизвестную *целевую зависимость* $y^*\colon X \to Y$ по обучающей выборке $X^\ell = \{x_1, \dots, x_\ell\}$, $y_i = y^*(x_i)$. Предполагается, что на множестве объектов определена функция расстояния $\rho(x, x')$.

Пусть $K\colon \mathbb{R}_+ \to \mathbb{R}_+$ — заданная невозрастающая функция.

Дипольным классификатором (или просто *диполем*) с *ядром* K, *полюсами* r^-, r^+ и *радиусом* d назовём функцию $a\colon X \to Y$ вида

$$a(x) = \operatorname{sign}\left(K\left(\frac{\rho(x, r^+)}{d}\right) - K\left(\frac{\rho(x, r^-)}{d}\right)\right). \qquad (1)$$

Диполь совпадает с байесовским классификатором, если предположить, что плотности классов имеют вид $p(x \mid y) = p(x; r^y, d) = \frac{1}{K_0} K\left(\frac{\rho(x, r^y)}{d}\right)$, где K_0 — нормировочный коэффициент; r^-, r^+, d — параметры диполя.

Введём также неотрицательную *функцию компетентности* диполя $C(x; r^-, r^+, R)$, значение которой непрерывно и монотонно убывает по мере удаления от полюсов диполя, а скорость убывания задаётся *радиусом* функции компетентности R, вообще говоря, отличным от радиуса диполя d. Будем рассматривать только такие функции компетентности, которые симметричны относительно полюсов диполя. Существует три принципиально различных варианта задания таких функций:

$$C(x;r^-,r^+,R) = K'\left(\tfrac{\rho(x,r^+)}{R}\right) K'\left(\tfrac{\rho(x,r^-)}{R}\right); \tag{2}$$

$$C(x;r^-,r^+,R) = K'\left(\tfrac{\rho(x,r^+)+\rho(x,r^-)}{2R}\right); \tag{3}$$

$$C(x;r^-,r^+,R) = K'\left(\tfrac{1}{R}\rho\bigl(x,\tfrac{r^+ + r^-}{2}\bigr)\right); \tag{4}$$

где $K'\colon \mathbb{R}_+ \to \mathbb{R}_+$ — заданная невозрастающая функция, вообще говоря, отличная от ядра K. Три варианта соответствуют трём способам агрегирования расстояний до полюсов: на уровне ядер (2), на уровне функций расстояния (3), и на уровне самих полюсов (4).

Композиция дипольных классификаторов

Композицией из T диполей с параметрами $\Theta = \bigl(r_t^-, r_t^+, d_t, R_t, Q_t\bigr)_{t=1}^T$, будем называть алгоритм классификации вида

$$a(x;\Theta) = \operatorname{sign}\left(\sum_{t=1}^T g_t(x,\Theta)\Bigl(p(x;r_t^+,d_t) - p(x;r_t^-,d_t)\Bigr)\right), \tag{5}$$

где $g_t(x,\Theta)$ — *шлюзовая функция* (gate [3]), оценивающая компетентность t-го диполя в точке x. Если шлюзовые функции нормированы,

$$g_t(x;\Theta) = \frac{Q_t C(x;r_t^+,r_t^-,R_t)}{\sum_{s=1}^T Q_s C(x;r_s^+,r_s^-,R_s)},$$

то композицию диполей можно интерпретировать как *смесь экспертов* [3], а сами шлюзовые функции — как априорные вероятности того, что объект x порождён t-й компонентой смеси. Возвращаясь к физической аналогии, параметр Q_t можно интерпретировать как заряд диполя.

Построение смеси диполей

Композицию диполей удобно настраивать с помощью EM-алгоритма. На каждой EM-итерации сначала вычисляются «скрытые переменные» — апостериорные вероятности h_{ti} того, что объект x_i порождён t-й компонентой смеси (E-шаг). Благодаря скрытым переменным задача максимизации правдоподобия распадается на T независимых подзадач, по одной для каждого диполя (M-шаг), которые решаются стандартными методами. Возможен вариант, когда полюса диполей выбираются только из обучающих объектов. Если правдоподобие значительной доли объектов оказывается слишком низким, и среди них есть достаточное количество объектов обоих классов, то в композицию вводится новый диполь.

В докладе подробно рассматриваются различные альтернативные варианты построения дипольной композиции, а также вопросы ускорения сходимости. В экспериментах на модельных данных алгоритм показал хорошее качество классификации линейно разделимой выборки и выборки типа XOR, построив один и два диполя соответственно.

Достоинства дипольных композиций

Благодаря вероятностной модели, наряду с классификацией, могут быть выданы оценки принадлежности объекта каждому из классов, а также выделены *нетипичные объекты*, находящиеся в зоне низкой компетентности всех диполей.

Классификации легко интерпретируются в терминах сходства: «объект x отнесён к классу y потому, что он близок к эталонному объекту x_i класса y». Эталонными объектами являются полюса диполей.

В зависимости от целей применения в роли полюсов могут выступать либо обучающие объекты, либо произвольные точки пространства X.

Благодаря EM-алгоритму с динамическим добавлением компонент строится минимальное необходимое число диполей, в отличие от таких композиционных алгоритмов, как бустинг и бэггинг.

Для хранения алгоритма достаточно запомнить только полюса диполей (и ещё три числа на каждый диполь). В этом дипольная композиция аналогична методу опорных векторов, SVM [5]. Однако, в отличие от SVM, *опорными* являются не пограничные обучающие объекты (которые часто оказываются шумовыми выбросами), а полюса диполей, найденные оптимизационной процедурой. В этом дипольная композиция аналогична методу релевантных векторов, RVM [4]. Однако, в отличие от RVM, здесь не делается никаких априорных предположений о виде вероятностного распределения в пространстве параметров модели.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00877 и № 05-07-90410, а также программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — P. 320.

[2] *Воронцов К. В.* О композициях дипольных классификаторов // Интеллектуализация обработки информации. — Симф., 2006. — С. 38–40.

[3] *Jordan M. I., Jacobs R. A.* Hierarchical mixtures of experts and the EM algorithm // Neural Computation. — 1994. — no. 6. — Pp. 181–214.

[4] *Tipping M.* The relevance vector machine // Advances in Neural Information Processing Systems, San Mateo, CA. — Morgan Kaufmann, 2000.

[5] *Vapnik V.* Statistical Learning Theory. — Wiley, New York, 1998.