Двухфазная схема решения задачи классификации

Евгений Прохоров, Михаил Кумсков
Механико-математический факультет
Московский государственный университет имени М.В. Ломоносова
Москва, Россия
eugeny.prokhorov@gmail.com, qsar_msu@mail.ru

Аннотация

Данная работа посвящена решению задачи классификации в общем случае. Результаты получены в терминах математической теории распознавания образов. Предложенные методы решают задачу построения правил отказа от прогноза при просмотре больших баз объектов. Для предложенной в работе двухфазной схемы доказана оценка качества прогнозирования. В частности доказано улучшение качества прогноза при использовании нетривиальных правил отказа от прогноза. Также приводятся результаты практических испытаний предложенного метода решения задачи, подтверждающие эффективность описанного подхода. Полученные результаты могут быть использованы для улучшения качества классификации изображений с помощью распознающих моделей различной природы, а также для эффективного поиска нужных изображений в больших банках данных.

Ключевые слова: *классификация, распознавание образов, машинное обучение*

1. ВВЕДЕНИЕ

Работа посвящена в первую очередь задаче бинарной классификации, когда необходимо с использованием ранее обученной модели отнести объект классификации к одной из двух групп (обычно обладающих и не обладающих ключевым свойством объектов). Далее в тексте эти две группы условно обозначаются как «активные» и «неактивные» объекты. Однако полученные результаты легко обобщаются и на задачу с произвольным конечным числом классов. Ключевой особенностью решаемой задачи распознавания является предположение о том, что построенные модели будут использоваться для поиска объектов, потенциально обладающих ключевым свойством, в больших базах данных. Примерами таких задач могут выступать виртуальный скрининг [7] (поиск потенциально активных химических веществ в больших базах соединений) или поиск изображений, сходных с заданным шаблоном.

С учетом этой особенности рассматриваются ограничения допустимости для распознающих моделей [3, 9]. Такой подход к решению задачи классификации может быть рассмотрен, как классификация с отказами [2] или как частный случай применения смесей экспертов для классификации [8].

2. ОПРЕДЕЛЕНИЯ И ПОСТАНОВКА ЗАДАЧИ

Пусть обучающая выборка LS состоит из N объектов x_i, $i = 1,\ldots,N$, каждому из которых поставлено в соответствие одно из значений: 1 или –1 (1 соответствует условно активным объектам, –1 – неактивным). Вектор, последовательно содержащий активности всех объектов обучающей выборки, обозначим $y = (y_1, y_2, \ldots, y_N)$, $y_i \in \{-1,1\}$.

Пусть также построена распознающая модель, решающая исходную задачу классификации, т.е. $RM_1(x_i) \in \{-1,1\}$ для любых $x_i \in LS$. Назовем RM_1 моделью первого уровня.

Напомним, что процедура скользящего контроля (leave-one-out cross-validation [4]) заключается в следующем: из обучающей выборки последовательно удаляется каждый объект, по оставшимся объектам строится распознающая модель, и с помощью этой модели прогнозируется активность удаленного объекта. В работе везде будет использован функционал качества моделей со скользящим контролем, равный отношению количества верных прогнозов к общему числу спрогнозированных объектов.

Обозначим через R_1 – множество тех объектов обучающей выборки x_i, для которых полученные в ходе процедуры скользящего контроля значения активности совпадают с действительными: $RM_1(x_i) = y_i$, т.е. множество верно классифицированных моделью первого уровня объектов. Через W_1 обозначим множество ошибочно классифицированных моделью первого уровня объектов: $W_1 = \{x_i \in LS \mid RM_1(x_i) \neq y_i\}$. Таким образом, функционал качества со скользящим контролем для модели первого уровня равен $\varphi_1 = |R_1|/N$.

Определим задачу классификации второго уровня. Всем объектам обучающей выборки, спрогнозированным верно моделью первого уровня (их $|R_1|$), поставим в соответствие значение 1, а соединениям, спрогнозированным неверно (их $|W_1|$), поставим в соответствие значение –1. Сформируем таким образом вектор $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$, $\hat{y}_i \in \{-1,1\}$:

$$\hat{y}_i = \begin{cases} 1, & \textit{если} \quad RM_1(x_i) = y_i; \\ -1, & \textit{если} \quad RM_1(x_i) \neq y_i, \end{cases} \quad i = 1,\ldots,N.$$

Возникшую задачу классификации назовем задачей классификации второго уровня.

Пусть построена распознающая модель RM_2, решающая задачу классификации второго уровня, т.е. $RM_2(x_i) \in \{-1,1\}$ для любых $x_i \in LS$. Назовем RM_2 моделью второго уровня.

Пусть в ходе процедуры скользящего контроля моделью второго уровня получено $|R_2|$ верных прогнозов, где $R_2 = \{x_i \in LS \mid RM_2(x_i) = \hat{y}_i\}$. Тогда функционал качества модели второго уровня $\varphi_2 = |R_2|/N$.

Наконец, определим результирующую распознающую модель RM_0. Результирующая модель решает исходную задачу классификации, но в отличие от модели первого уровня результирующая модель обладает опцией отказа от прогноза. То есть $RM_0(x_i) \in \{-1,0,1\}$ $\forall x_i \in LS$ и значение $RM_0(x_i) = 0$ интерпретируется как отказ от прогноза активности объекта x_i.

Для $x_i \in LS$

$$RM_0(x_i) = \begin{cases} 1, & \textit{если} \quad RM_2(x_i) = 1 \quad \textit{и} \quad RM_1(x_i) = 1; \\ -1, & \textit{если} \quad RM_2(x_i) = 1 \quad \textit{и} \quad RM_1(x_i) = -1; \\ 0, & \textit{если} \quad RM_2(x_i) = -1. \end{cases}$$

Таким образом, результирующая модель осуществляет отказ от прогноза тогда, когда модель второго уровня предсказывает, что модель первого уровня ошибается, и осуществляет прогноз активности моделью первого уровня в противном случае.

Как и ранее, обозначим через $R_0 = \{x_i \in LS \mid RM_0(x_i) = y_i\}$ множество верно классифицированных результирующей моделью объектов. Пусть также через $\mathrm{R}eject$ обозначено количество отказов от прогноза. Тогда функционал качества результирующей модели $\varphi_0 = |R_0|/(N - \mathrm{R}eject)$.

3. ОЦЕНКА КАЧЕСТВА РЕЗУЛЬТИРУЮЩЕЙ МОДЕЛИ

Теорема. Верна следующая оценка качества результирующей модели.

$$\varphi_0 = \frac{(\varphi_1 + \varphi_2)N - \mathrm{R}eject}{2(N - \mathrm{R}eject)}.$$

Доказательство. По пределению $\varphi_1 N = |R_1|$, а $\varphi_2 N = |R_2|$. Кроме того $R_0 = R_1 \cap R_2$. Последнее можно доказать так: если $x_i \in R_1 \setminus R_2$, то модель первого уровня осуществляет верный прогноз, однако модель второго уровня (ошибаясь) возвращает значение –1, в силу чего $RM_0(x_i) = 0$, таким образом, происходит отказ от прогноза. Если же $x_i \in R_2 \setminus R_1$, то модель первого уровня ошибается, а модель второго уровня снова возвращает –1, что опять означает отказ от прогноза. Когда $x_i \notin R_1$ и $x_i \notin R_2$, модель первого уровня ошибается, в то время как модель второго уровня возвращает значение 1, таким образом, осуществляется неверное прогнозирование.

Учитывая выше сказанное, заметим, что отказам от прогноза соответствуют множества $R_2 \setminus R_1$ и $R_1 \setminus R_2$. Следовательно, $\mathrm{R}eject = |R_1 \Delta R_2|$.

Таким образом, числитель дроби в формулировке теоремы приобретает вид $|R_1| + |R_2| - |R_1 \Delta R_2|$. Далее,

$$|R_1| + |R_2| - |R_1 \Delta R_2| = 2|R_1 \cap R_2| = 2|R_0|,$$

и, сокращая дробь на 2, имеем выражение для функционала качества φ_0. W

Следствие 1. Пусть $\varphi_{\min} = \min(\varphi_1, \varphi_2) > 1/2$, тогда, если $\mathrm{R}eject > 0$, то $\varphi_0 > \varphi_{\min}$.

Доказательство. Заметим, что $\mathrm{R}eject < N$. Действительно, так как $\varphi_1 > 1/2$ и $\varphi_2 > 1/2$, то $R_1 \cap R_2 \neq \varnothing$ и существуют верные ответы результирующей модели.

Теперь по теореме имеем

$$\varphi_0 = \frac{(\varphi_1 + \varphi_2)N - \mathrm{R}eject}{2(N - \mathrm{R}eject)}, \Rightarrow \varphi_0 \geq \frac{2\varphi_{\min} N - \mathrm{R}eject}{2(N - \mathrm{R}eject)} =$$

$$= \frac{2\varphi_{\min} N - 2\varphi_{\min} \mathrm{R}eject + 2\varphi_{\min} \mathrm{R}eject - \mathrm{R}eject}{2(N - \mathrm{R}eject)} =$$

$$= \frac{2\varphi_{\min}(N - \mathrm{R}eject) + (2\varphi_{\min} - 1)\mathrm{R}eject}{2(N - \mathrm{R}eject)} =$$

$$= \varphi_{\min} + \frac{(\varphi_{\min} - 1/2)\mathrm{R}eject}{N - \mathrm{R}eject} > \varphi_{\min}.$$

W

Следствие 2. Если $\varphi_2 \geq \varphi_1 > 1/2$, то в случае $\mathrm{R}eject > 0$ имеем $\varphi_0 > \varphi_1$.

Доказательство тривиально вытекает из следствия 1. W

Следствие 3. Если $\varphi_2 > \varphi_1 > 1/2$, то $\varphi_0 > \varphi_1$.

Доказательство вытекает из доказателдьства следствия 1. W

Таким образом, доказано, что если модель первого уровня классифицировала объекты из обучающей выборки хотя бы чуть лучше, чем случайным образом, и качество модели второго уровня не хуже качества модели первого уровня, то при условии, что количество отказов от прогноза больше нуля, результирующая модель демонстрирует более высокое качество классификации на исходной задаче, чем модель первого уровня. Также, доказано улучшение качества результирующего прогноза в случае, когда качество модели второго уровня превосходит качество модели первого уровня.

4. ОПИСАНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

Описанный в работе подход тестировался на практике для решения задачи «структура – свойство». Задача «структура – свойство» – актуальная задача приложения математической теории распознавания образов в химии. Построенные классифицирующие модели используются для прогнозирования активности химических соединений. Работа с изображениями также является актуальной облатсью приложения предложенных методов, однако, на текущий момент тестирование на изображениях не проводилось и является направлением будующей работы.

В таблице приведены результаты использования предложенной двухфазной схемы решения для прогнозирования активности ингибиторов фермента деления клеток (PARP) [1]. Обучающая выборка, предоставленная Высшим химическим колледжем РАН, состояла из 120 соединений. Экспериментальное измерение активности проводилось в Институте химической биологии и фундаментальной медицины СО РАН.

В качестве метода построения распознающих моделей первого и второго уровня выступал метод опорных векторов (Support Vector Machine) [6]. При его использовании применялось ядро многослойного персептрона (Multilayer Perceptron kernel) [5]. Для задач классификации первого и второго уровня независимо применялся также эволюционный отбор признаков. В таблице столбцы D_1 и D_2 содержат количество признаков, отобранных для решения задач классификации первого и второго уровня соответственно. Строки таблицы соответсвуют различным типам описания химической структуры соединений (различным пространствам признаков).

Признаки	φ_1	φ_2	Отказы	φ_0	D_1	D_2
1	0,942	1	7	1	2	1
2	0,892	0,883	1	0,891	5	1
3	0,908	0,933	13	0,972	3	3
4	0,942	0,925	9	0,982	4	1
5	0,883	0,892	9	0,919	1	1
6	0,875	0,875	8	0,911	3	1
7	0,908	0,925	8	0,946	2	1
8	0,967	0,925	9	0,982	4	1

Таблица 1. Качество прогноза моделей, построенных с помощью метода опорных векторов

На основе полученных результатов построен прогноз для соединений с неизвестной активностью. Для наиболее перспективных соединений проводилось экспериментальное оценивание ингибирующей концентрации, в результате чего было найдено как минимум одно активное соединение, являющееся ингибитором PARP. Таким образом, показана практическая применимость двухфазной схемы решения задачи классификации.

5. ЗАКЛЮЧЕНИЕ

В работе предложен новый подход к решению задачи классификации, учитывающий такую особенность некоторых прикладных задач классификации, как необходимость осуществлять классификацию только ограниченного числа объектов. Подход может быть применен для решения задач виртуального скрининга и для поиска изображений, соответствующих заданному шаблону. Доказаны теоретические оценки качества прогнозирования при использовании описанных методов. Приведенные практические исследования показывают эффективность подхода. Тестирование подхода на больших базах изображений является направлением будущей работы.

6. БЛАГОДАРНОСТИ

Работа была выполнена при частичной поддержке грантов РФФИ №10-07-00694, №12-03-01036-а и №12-03-92420.

7. ССЫЛКИ

[1] Amours D., Desnoyers S., Silva I. and Poirier G.G. Poly(ADP-ribosyl)ation reactions in the regulation of nuclear functions // Biochem. J. 1999. **342**, № 2. 249–268.

[2] Herbei R., Wegkamp M. Classification with reject option // Can. J. Statist. 2006. **4**, №4. 709–721.

[3] Prokhorov E.I., Ponomareva L.A., Permyakov E.A., Kumskov M.I. Fuzzy classification and fast rules for refusal in the QSAR problem // Pattern Recogn. and Image Anal. 2011. **21**, №3. 542–544.

[4] Stone M. Cross-validatory choice and assessment of statistical predictions. // J. Roy. Statist. Soc. 1974. **B**, №36. 111–147.

[5] Thomas R., Karsten B. Multilayer Perceptron kernel // Proc. 24th SIBGRAPI Conf. on Graphics, Patterns and Images. Maceio, Alagoas, Brazil, 2011. 337–343.

[6] Vapnik V.N. The nature of statistical learning theory. N. Y.; London: Springer, 1998.

[7] Walters W.P., Stahl M.T., Murcko M.A. Virtual screening – an overview // Drug Disc. Today 1998. V. 3. P. 160-178.

[8] Yuksel S.E., Wilson J.N., Gader P.D. Twenty years of mixture of experts. // IEEE Trans. Neural Networks Learning Syst. 2012. **23**, №8. 1177–1193.

[9] Прохоров Е.И. Нейронные сети для построения ограничений допустимости в задаче «структура–свойство» // Нейрокомпьютеры: разработка, применение. 2012. №10. 46–56.

Об авторах

Евгений Прохоров – аспирант кафедры вычислительной математики механико-математического факультета МГУ им. Ломоносова. Электронная почта: eugeny.prokhorov@gmail.com

Михаил Кумсков – доктор физико-математических наук, профессор механико-математического факультета МГУ им. Ломоносова. Электронная почта: qsar_msu@mail.ru