РЕГУЛЯРИЗАЦИЯ МЕТОДА РЕШАЮЩИХ ДЕРЕВЬЕВ, ОСНОВАННАЯ НА ПРИНЦИПЕ УСТОЙЧИВОСТИ

Д. Ветров[1], Д. Кропотов[2], И. Толстов[3]

[1] 119991, Москва, ул.Вавилова 40, ВЦ им. А.А.Дородницына РАН, VetrovD@yandex.ru
[2] 119991, Москва, ул.Вавилова 40, ВЦ им. А.А.Дородницына РАН, DKropotov@yandex.ru
[3] 119234, Москва, Ленинские горы, МГУ, 2ой учебный корп., igor_tolstov@mail.ru

В работе рассматривается новый подход к построению бинарного решающего дерева, обеспечивающего наименьший процент ошибок на независимой выборке. Предложена концепция регуляризации функционала качества, препятствующая излишней настройке алгоритма на обучающую выборку. Выводы работы подтверждены результатами большого числа экспериментов.

Введение

Методы построения бинарных решающих деревьев [1],[6] являются одними из наиболее широко используемых методов распознавания образов. Их достоинствами являются быстрое время работы, нелинейный характер получаемой разделяющей поверхности. Одной из наиболее распространенных концепций обучения являяются т.н. методы обрезания (pruning), т.е. замены части поддеревьев листьями [3],[5],[7]. На первом этапе строится полное дерево, безошибочно классифицирующее все объекты обучающей выборки. Затем производится уменьшение размерности полученного дерева с целью улучшения его обобщающей способности. Большинство методов обрезания используют различные эвристики, направленные на максимальное уменьшение размерности дерева с минимальной потерей качества распознавания обучающей выборки. В работе [2] было показано, что использование независимой выборки для контроля обобщающей способности при обрезании, не приводит к улучшению качества по сравнению с обычными методами обрезания. Таким образом, можно заключить, что концепция эвристического обрезания решающих деревьев достигла максимума своей эффективности.

В настоящей работе предлагается иной подход к обрезанию дерева, опирающийся на косвенные оценки способности модели к обобщению. Добавление к функционалу качества члена, отвечающего за обобщение (регуляризация) приводит к новому правилу обрезания, основанному на минимизации регуляризованного функционала.

Дальнейшее изложение материала построено следующим образом. Во второй главе дана концептуальная схема процесса восстановления зависимостей по конечным наборам данных, которая, после некоторых модификаций, рассматривается в главе 3 применительно к методам построения бинарных решающих деревьев. В этом случае, в качестве искомого будет использоваться дерево, при котором достигается минимум некоторого функционала. Подвергая классический функционал качества регуляризации, мы получаем семейство «наилучших» (с точки зрения функционала) деревьев. Вопросы выбора наилучшего параметра регуляризации рассматриваются в главе 4. В пятой главе изложены результаты испытаний регуляризованных деревьев и сравнения качества их работы с известными методами обрезания на нескольких широко-распространенных задачах.

Запоминание и обобщение

Задача машинного обучения как задача нахождения скрытых зависимостей по конечным наборам данных предполагает одновременное решение двух разных проблем. Во-первых, необходимо найти закономерности, отвечающие за вид данной обучающей выборки. Во-вторых, нужно выделить из них те закономерности, которые влияют на вид генеральной совокупности. Именно они и должны использоваться для последующего анализа данных. Плохое решение первой задачи приводит к низкому качеству работы алгоритма даже на обучающих данных. Следовательно, в этом случае тем более нельзя ожидать хорошей работы на независимых выборках. Пренебрежение второй задачей приводит к перенастройке на обучающую выборку и вырождению алгоритма при обработке новой информации.

Применительно к задаче распознавания образов, решение первой проблемы означает увеличение запоминающей способности. Она успешно решается классическими методами оптимизации посредством максимизации доли правильно распознанных объектов обучающей выборки. Решение второй задачи означает увеличение обобщающей способности, которая может быть выражена, например, следующей формулой:

$$GC = 1 - P_{train} + P_{univ}$$

где P_{train} - доля ошибок на обучающей выборке, а P_{univ} - вероятность ошибочной классификации объекта генеральной совокупности. Последний показатель недоступен прямому измерению, поэтому вместо него используются его оценки (полученные, например, с помощью скользящего контроля или независимой валидационной выборки). Эти оценки обычно требуют значительных временных или информационных затрат. Другим способом может являться использование некоторых косвенных характеристик обобщающей способности. Таковым может являться следующий принцип устойчивости: *обобщающая способность тем выше, чем дальше, в среднем, расположены правильно классифицированные объекты от предположительной границы класса*.

Заметим, что в алгоритмах обучения бинарных решающих деревьев этот принцип не учитывается даже косвенно, а значит, мы можем использовать обучающую выборку для определения среднего расстояния между объектом и границей класса. Вообще, условие устойчивости близко к принципу «максимального зазора», использующемуся, например, в методе опорных векторов [8].

Регуляризация функционала

В существующих алгоритмах обучения построение решающего дерева заканчивается по достижении нулевого уровня ошибки на обучающей выборке. Другими словами, на первом этапе обучения минимизируется следующий функционал

$$\Phi_0 = P_{train}$$

Введем поправку, которая бы контролировала обобщающую способность. В дальнейшем будем полагать, что все признаки соответствующим образом отмасштабированы и имеют единичную дисперсию. Обозначим C_p множество правильно классифицированных объектов обучающей выборки. Рассмотрим следующий функционал качества

$$\Phi_\lambda = P_{train} + R(\lambda) \qquad (1)$$

где

$$R(\lambda) = \frac{1}{mP_{train}} \sum_{S \in C_p} \exp\left(-\frac{\rho^2(S, \partial K)}{\lambda^2} \right)$$

В последней формуле $\rho(S, \partial K)$ - расстояние от объекта до границы листа, в котором оказался объект, m - количество объектов обучающей выборки, а λ - параметр регуляризации. Дерево, на котором в ходе процесса обучения достигается минимум функционала (1), будет использоваться для последующего распознавания. Очевидно, имеет место следующая теорема

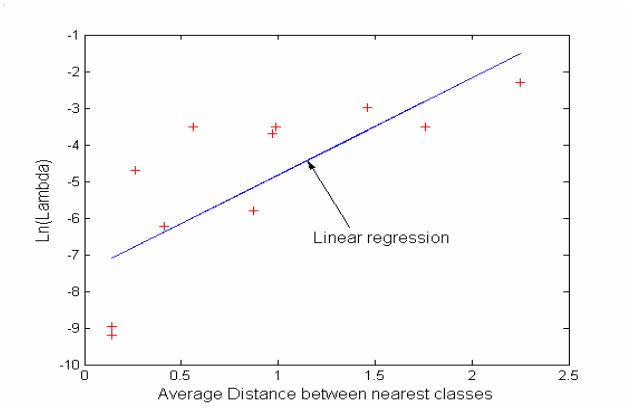

Рисунок. Уравнение регрессии для параметра регуляризации.

Теорема *(о сходимости)***.** При стремлении параметра регуляризации к нулю, функционал (1) стремится к доле ошибок на обучении

$$\lim_{\lambda \to 0} \Phi_\lambda = \Phi_0$$

Следствие. Для любой задачи распознавания образов найдется такое $\varepsilon > 0$, что при всех $\lambda < \varepsilon$ минимизация регуляризованного и исходного функционалов приведут к построению одинакового дерева.

Нетрудно заметить, что оба слагаемых в регуляризованном функционале принимают значения от нуля до единицы, причем в процессе ветвления первое слагаемое убывает, а второе возрастает. При надлежащем выборе параметра регуляризации функционал Φ_λ будет иметь минимум, а соответствующее дерево будет близко к оптимальному (т.е. допускающему минимум ошибок на объектах генеральной совокупности).

Выбор параметра регуляризации

Многочисленные испытания, проведенные на модельных и практических задачах, выявили связь (см. рис.) наилучшего параметра регуляризации (его качество оценивалось по доле ошибок на независимой тестовой выборке) и среднего расстояния между ближайшими классами, вычисляемого по формуле

$$x = \frac{1}{l} \sum_{i=1}^{l} \min_{j \neq i} \rho(K_i, K_j)$$

где l - число классов. Уравнение соответствующей регрессии имеет вид

$$\lambda = \exp(2.64x - 7.46)$$

Другим возможным способом определения наилучшего λ является разбиение обучающей выборки на предобучающую и валидационную. Последняя используется для оценки λ. Затем обе выборки снова объединяются в одну, а найденный параметр регуляризации используется в обучении по объединенной выборке. Эксперименты показали, что значения наилучшего λ для предобучающей и собственно обучающей выборок практически не отличаются.

Результаты экспериментов

Для сравнения качества работы регуляризованных деревьев были использованы методы и результаты статьи [2]. Необходимые данные были взяты из доступных источников (UCI repository)[4], содержащих стандартные прикладные задачи, использующиеся для сравнения различных методов распознавания. Для каждой задачи были сгенерированы случайным образом 25 обучающих выборок, содержащих 70% объектов, а из оставшихся объектов формировались тестовые выборки. Результаты распознавания усреднялись для каждой задачи, а полученный процент сравнивался с соответствующими результатами, полученными после применения различных методов обрезания. Для определения параметра регуляризации производилось предварительное разбиение обучающей выборки на первой итерации. Найденное значение λ использовалось в остальных 24 итерациях. Итоги сравнения приведены в таблице. В первом столбце показан процент ошибок при использовании полного дерева. В следующих пяти столбцах даны результаты работы пяти различных методов обрезания (подробнее см. [2]), а в последнем столбце указан результат работы регуляризованного дерева. Эти, а также серия других экспериментов позволяют

Таблица. Результаты распознавания различными методами обрезания.

	Full Tree	REP	MEP	CVP	PEP	EBP	R-Tree
Iris	6.2	5.68	***6.22***	5.87	5.33	5.07	**4.93**
Glass	35.38	***38.5***	38.2	36.87	**35.31**	35.88	35.89
Cleveland	29.1	27.65	28.88	***30.07***	29.01	28.88	**23.25**
Switzerland	13.3	6.27	***12.65***	**2.39**	6.16	6.16	9.72
Pima	31.43	**25.88**	27.22	30	***28.85***	28.84	26.29
Australian	18.71	15.13	15.07	***18.47***	**14.59**	15.42	15.76

сделать вывод о том, что для ряда задач регуляризация процесса обучения является более предпочтительной по сравнению с обрезанием. Регуляризатор $R(\lambda)$, в целом, может служить косвенным показателем обобщающей способности при надлежащем выборе параметра λ. Само значение параметра регуляризации является специфическим для каждого конкретного случая, зависящим скорее от топологии генеральной совокупности (в частности, обучающей выборки), чем от размерности задачи.

Заключение

В данной работе предложен новый подход к процедуре обучения бинарных решающих деревьев, основанный на регуляризации функционала качества. Благодаря тому, что значение регуляризатора $R(\lambda)$ никак не влияет на процесс обучения, можно проводить его подсчет по обучающей выборке и использовать его значение в качестве косвенной характеристики степени перенастройки на выборку. Сравнение с существующими методами обрезания, использующимися для упрощения полных деревьев, показывает, что описанный подход, в целом, не уступает, а в ряде случаев превосходит их. Основной трудностью является выбор наилучшего параметра регуляризации. Одним из направлений дальнейшей работы будет поиск взаимосвязей между топологией задачи и наилучшим (или близким к нему) значением λ. Заметим, что подобная регуляризация может быть использована не только для метода решающих деревьев, но и во многих других семействах алгоритмов распознавания образов.

Работа выполнена при частичной поддержке РФФИ (гранты 02-07-90134, 02-07-90137, 03-01-00580, 02-01-00558 и 04-01-00161).

Список литературы

1. L. Breiman, J. Friedman, R. Olshen, C. Stone. Classification and Regression Trees. Belmont, Calif.: Wadsworth Int'l, 1984.
2. F. Esposito, D. Malerba, G. Semeraro. A Comparative Analisys of Methods for Pruning Decision Trees // IEEE Transactions on Pattern Analisys and Machine Intelligence, Vol. 19, No. 5, pp. 476-492, 1997.
3. J. Mingers. Expert Systems – Rule Induction With Statistical Data // Operational Research Society, Vol. 38, pp. 39-47, 1987.
4. P.M. Murphy, D.W. Aha. UCI Repository of Machine Learning Databases [Machine Readable Data Repository], Univ. of California, Dept. of Information and Computer Science, Irvine, Calif., 1996.
5. T. Niblett. Constructing Decision Trees in Noisy Domains // Progress in Machine Learning, Proc. EWSL 87, Wilmslow: Sigma Press, pp. 67-78, 1987.
6. J.R. Quinlan. Induction of Decision Trees // Machine Learning, Vol. 1, No. 1, pp. 81-106, 1986.
7. J.R. Quinlan. C4.5: Programs for Machine Learning. San Mateo, Calif.: Morgan Kaufamnn, 1993.
8. V. Vapnik. Statistical Learning Theory. New York, Wiley, 1998.