УДК 519.68

СТРАТЕГИИ РЕДУКЦИИ РЕШАЮЩИХ ДЕРЕВЬЕВ (ОБЗОР)

Ю.Ю. Дюличева

Таврический национальный университет, факультет математики и информатики,
ул. Ялтинская, 4, Симферополь, Украина,
e-mail: *dyulicheva@fromru.com*

Abstract

Some pruning strategies of decision trees that to overfitting avoidance on training set and their negative and positive properties are considered in the article. The lack of theoretical justification for existent pruning strategies of decision trees is noted.

Введение

На протяжении десятилетий усилия исследователей в области обучения распознаванию на основе решающих деревьев (РД) [1, 3, 6, 9, 11] были направлены на разработку стратегий избежания переподгонки на обучающем множестве, основными из которых являются стратегии редукции решающих деревьев.

Многие современные алгоритмы синтеза РД по обучающей выборке включают три стадии: синтез РД, редукция РД и собственно распознавание.

Процесс избыточного усложнения решающего дерева, вызванный излишним следованием «зашумленным» данным, называется *переподгонкой* (*overfitting*). Результатом переподгонки на обучающем множестве является РД с излишне сложной структурой, отражающей не только основные закономерности в данных, но и влияние *случайных объектов*, что приводит к снижению эффективности процесса распознавания на объектах контрольной выборки. Процесс упрощения РД за счет отсечения избыточных ветвей с целью избежания переподгонки называется *редукцией (pruning)*. Суть редукции состоит в удалении таких поддеревьев РД, которые характеризуются минимальной статистической достоверностью. Однако процесс редукции не гарантирует улучшения средней (обобщенной) точности классификатора типа РД, хотя и влечет уменьшение числа листьев. Процесс редукции направлен на сокращение зависимости РД от «зашумленных» данных.

Предредукция (prepruning) и *постредукция (postpruning)* — две стандартные эвристические стратегии редукции решающих деревьев. Предредукция или критерий «ранней остановки» досрочно прекращает дальнейшее ветвление в вершине РД, основываясь на некоторой эвристической мере (например, используя информационный прирост, критерий χ^2 или точный критерий Фишера). Предредукция не является эффективным методом избежания переподгонки, поскольку принятие решения в каждой вершине РД осуществляется на основе локальной информации в этой вершине (не учитывая информацию о том, что произойдет на нижних уровнях дерева) и, как следствие, принятое решение лишь аппроксимирует оптимальное. Как правило, более эффективной считается стратегия постредукции. Стратегия постредукции осуществляет отсечение ветвей, согласно эвристической мере (например, на основе коэффициента ошибки в каждой вершине), после того, как дерево полностью «настроится» на имеющуюся обучающую выборку.

Рассмотрим наиболее известные стратегии постредукции. Традиционно выборка разбивается на обучающее (70% данных) и контрольное множество (30% данных).

1. Редукция на основе сокращенной ошибки (Reduced Error Pruning(REP), Quninlan, 1987) [2, 4, 5, 6]

Согласно стратегии *REP* обучающее множество разбивается на два подмножества: множество синтеза РД (70% от обучающего множества) и множество редукции РД (30% от обучающего множества). РД T полностью «настраивается» на объекты из множества синтеза РД. Синтез редуцированного дерева производится на объектах множества редукции. Для каждой внутренней вершины t осуществляется сравнение числа ошибок классификации на множестве редукции $I(T(t))$, допускаемых поддеревом с корневой вершиной t с числом ошибок классификации на множестве редукции $I(t)$, которое возникает в результате преобразования вершины t в лист согласно мажоритарному правилу. Если $I(T(t)) \geqslant I(t)$, то поддерево с корневой вершиной t редуцируют. Далее процесс редукции применяется к полученному редукцированному дереву до тех пор, пока для всехвнутренних вершин не будет выполнено $I(T(t)) < I(t)$. За счет того, что объекты множества редукции не участвуют в синтезе РД, будет получена несмещенная оценка ошибки классификации на объектах контрольного множества.

Стратегия *REP* не указывает, как выбирать метки класса для листьев, получаемых в процессе редукции. Здесь существует две возможности: использовать мажоритарное правило класса на объектах обучающей выборки или на объектах множества редукции.

Так как стратегия *REP* использует разные множества для синтеза и редукции РД, то в процессе редукции часто возникает ситуация, когда некоторые поддеревья РД не получают объекты из множества редукции. Такие поддеревья (пустые поддеревья) принято считать формируемыми за счет случайных объектов, попавших во множество синтеза и, следовательно, они всегда редуцируются стратегией *REP*. Пустые поддеревья связаны с исследованием интервалов наименьшей размерности, соответствующих ветви дерева и покрывающих сравнительное небольшое число объектов обучающего множества.

Пусть r_L - ошибка листа с меткой мажоритарного класса, который заменит в процессе редукции поддерево $T(t)$ с корнем в вершине t, тогда r_L зависит только от (распределения классов) объектов, которые достигают корневой вершины t поддерева $T(t)$. Другими словами, структура дерева, расположенная выше поддерева $T(t)$ определяет величину ошибки r_L. Пусть $r_{T(t)}$ - ошибка поддерева $T(t)$ после проведения редукции в этом поддереве согласно стратегии *REP*, тогда величина ошибки $r_{T(t)}$ либо уменьшилась либо осталась неизменной. Таким образом, r_L и $r_{T(t)}$ - независимые случайные величины. Следовательно, вероятность события $r_{T(t)} < r_L$ увеличивается в процессе редукции дерева $T(t)$. Это смещение в распространении ошибки свойственно стратегии *REP*. В работе [4] отмечается, что чем сложнее исходное дерево и меньше объектов во множестве редукции, тем сильнее проявляется этот эффект.

В результате реализации стратегий *REP* будет получено редуцированное РД с наибольшей точностью классификации на объектах множества редукции [4, 5]. *REP* имеет линейную вычислительную сложность по числу вершин, поскольку каждая вершина в процессе редукции просматривается ровно один раз. К недостаткам стратегии следует отнести ее склонность к чрезмерной редукции РД, особенно для данных, в которых число объектов множества редукции существенно меньше числа объектов множества синтеза РД.

2. Редукция на основе пессимистической ошибки (Pessimistic Error Pruning(PEP), Quinlan, 1987) [2, 5]

Стратегия использует одно и тоже обучающее множество и для синтеза и для редукции РД. Пусть коэффициент ошибки в вершине t на объектах обучающего множества

$$r(t) = \frac{e(t)}{n(t)},$$

где $e(t)$ - число объектов обучающего множества, достигших вершины t. Поскольку обучающее множество содержит «зашумленные» объекты, несмешанная оценка коэффициента ошибки в вершине t на объектах обучающего множества множества имеет вид:

$$r'(t) = \frac{e(t) + \frac{1}{2}}{n(t)}.$$

Аналогично, для поддерева $T(t)$ с вершиной t несмещенный коэффициент ошибки на обучающем множестве имеет вид:

$$r'(T(t)) = \frac{\sum\limits_{j \in L(T(t))} e(j) + \frac{|L(T(t))|}{2}}{\sum\limits_{j \in L(T(t))} n(j)},$$

где *L(T(t))* - множество листьев $T(t)$. Редукция поддерева осуществляется, если $e'(t) \leqslant e'(T(t)) + SE[e'(T(t))]$, где

$$SE[e'(T(t))] = \left[e'(T(t)) \cdot \frac{n(t) - e'(T(t))}{n(t)}\right]^{\frac{1}{2}} -$$

стандартная ошибка поддерева $T(t)$, вычисленная в предложении, что распределение ошибок биномиально. Стратегия *PEP* просматривает вершины в процессе редукции сверху вниз, начиная с корня. Если ветвь поддерева $T(t)$ редуцируется, то потомки вершины t не исследуются в процессе редукции.

Существенным недостатком этой стратегии является отсутствие теоретического обоснования введения несмещенной оценки ошибки на обучающем множестве. В работе [5] отмечается, что стратегия *PEP* может привести как к недостаточной, так и чрезмерной редукции РД.

3. Редукция, основанная на ошибке (Error Based Pruning(EBP), Quninlan, 1993) [2, 5]

Стратегия использует обучающее множество и для синтеза и для редукции РД. В процессе редукции просмотр вершин осуществляется снизу вверх, начиная с листьев. Множество объектов, покрываемое листом t будем рассматривать в качестве статистической выборки, тогда введем доверительный интервал $[L_{CF}(t), U_{CF}(t)]$ для оценки коэффициента ошибки в t на обучающем множестве. Верхняя граница интервала определяется следующим образом: обозначим $P(e(t)/n(t) \leqslant U_{CF}) = CF$, где CF - уровень значимости. В предложении, что ошибки обучающего множества распределены биномиально с вероятностью p в $n(t)$ испытаниях, можно вычислить точное значение U_{CF} как величину p, для которой биномиально распределенная случайная величина X имеет не более $e(t)$ благоприятных исходов в $n(t)$ испытаниях с вероятностью CF, т.е. $P(X \leqslant e(t)) = CF$. После нахождения верхней оценки коэффициенты ошибок листьев и поддеревьев вычисляются в предположении, что они используются для классификации объектов контрольного множества такой же мощности, что и обучающее множество. Таким образом, прогнозируемый коэффициент ошибки для t равен $n(t) \cdot U_{CF}$. Сумма прогнозируемых коэффициентов ошибок всех листьев ветви $T(t)$ рассматривается как оценка коэффициента ошибки этой ветви. Сравнивая прогнозируемый коэффициент ошибки t с ошибками ветви $T(t)$ и наибольшей из ветвей с корнем в дочерней вершине вершины t, принимается решение о том оставлять без изменений $T(t)$, редуцировать или наращивать в вершине t.

4. Редукция на основе минимальной цены-сложности (Minimum Cost-Complexity Pruning (CCP), Breiman, 1984) [2, 5]

Эта стратегия реализована в классической системе CART (Classification and Regression Trees) и включает две стадии:

1. Согласно некоторой эвристике, полностью разветвленное РД T порождает параметрическое семейство поддеревьев $\{T_0, T_1, \ldots, T_r\}$.
2. Выбор наилучшего дерева T_i производится на основе оценки коэффициентов ошибок деревьев из параметрического семейства.

На первой стадии вводится эвристическая мера $\alpha = (r(t) - r(T(t)))/(|L(T(t))| - 1)$, измеряющая прирост цены (доли неправильно расклассифицированных объектов) при замене поддерева листом и характеризующая сложность. Решающее дерево T_{i+1} получается путем редукции всех вершин в T_i с наименьшей величиной α (первое дерево T_0 получается редуцированием тех ветвей исходного дерева T, для которых $\alpha = 0$). Каждое дерево T_i характеризуется определенным значением α_i, для которого выполняется условие $\alpha_i < \alpha_{i+1}$. Следовательно, множество $\{T_0, T_1, \ldots, T_r\}$ является параметрическим семейством деревьев, которое будем обозначать $T(\alpha)$. Решающие деревья параметрического семейства, построенные согласно стратегии CCP являются вложенными, поскольку при последовательной редукции каждое дерево содержит все вершины следующего

(с меньшим числом листьев) дерева из последовательности. Поначалу при переходе от очередного дерева к последующему редуцируется, как правило, большое число вершин, однако, по мере приближения к корню на каждом шаге будет редуцироваться все меньше вершин. Деревья параметрического семейства редуцируются оптимально в том смысле, что каждое дерево из семейства имеет меньшую цену среди всех деревьев с таким же числом листьев. Параметрическое семейство может быть построено за время равное квадрату числа внутренних вершин.

На второй стадии выбирается решающее дерево с наибольшей точностью из параметрического семейства $T(\alpha)$. Авторы *CCP* предлагают два способа оценки коэффициента ошибки каждого РД из $T(\alpha)$:

- использование множеств кросс-проверки;
- использование независимого множества редукции.

К сожалению, отсутствует какое-либо теоретическое обоснование стратегии *CCP* и, как показано в [5], стратегия *CCP* склонна к недостаточной редукции РД.

5. Редукция, основанная на минимальной ошибке (Minimum Error Pruning (MEP), Niblett & Bratko, 1986) [8]

Стратегия *MEP* использует одно и тоже обучающее множество и для синтеза и для редукции РД, осуществляя редукцию восходящим способом. В качестве эвристической меры используется:

$$E_{\omega_k}(t) = \frac{n(t) - n_{\omega_k}(t) + \ell - 1}{n(t) + \ell}, \tag{1}$$

где n_{ω_k} - число объектов мажоритарного класса ω_k в вершине t, ℓ — число классов. Если оценка в вершине t с редукцией меньше, чем без нее, то поддерево редуцируется, и корень этого поддерева заменяется листом.

В качестве недостатка можно отметить необоснованность названия $E_{\omega_k}(t)$ оценкой вероятности ошибки в вершине. С точки зрения частотных оценок, обоснованным было бы использование оценки

$$E^*_{\omega_k} = \frac{n(t) - n_{\omega_k}(t)}{n(t)} = 1 - \nu_{\omega_k}(t)$$

как частоты попадания в интервал, соответствующий вершине t объектов немажоритарного класса. Но в этом случае редукция, которая может рассматриваться как «стягивание» некоторых интервалов в один интервал большей размерности, не представляется целесообразной при большом числе классов ℓ. Понятие мажоритарного класса может вообще «потеряться».

Поэтому эвристическая оценка (1) представляется целесообразной с модификацией

$$E'_{\omega_k}(t) = \frac{n(t) - n_{\omega_k}(t) - \ell(t) - 1}{n(t) + \ell(t)},$$

где $\ell(t)$ - число классов, достижимых из вершины t; $\ell(t) \leqslant \ell$.

В работе [12] отмечается, что в случаях, когда при синтезе РД используется стратегия избежания переподгонки, возможно получение индукторов с большей ошибкой, чем при использовании более тщательно настроенных на обучающую выборку деревьев. К сожалению, в этой работе отсутствуют строгие обоснования выводов, касающихся переподгонки. Кривая в координатах «ошибка» (p) — «сложность РД» (μ-число листьев) [12] (рис.1) верно характеризует процесс синтеза. Действительно, при малом $\mu \ll \mu_{\text{опт}}$ появляется все больше ошибок на обучающей выборке (и, как следствие, при распознавании произвольных объектов). При $\mu \gg \mu_{\text{опт}}$, $\mu \approx m_0$, $\mu \leqslant m_0$, получается «вырожденный» классификатор, настроенный «поточечно».

Используя свойства единичных интервалов для почти всех булевых функций можно обосновать положения работы [12]. Нужно отметить, что принципиальное значение имеет *модель начальной информации*: являются ли данные таблицы обучения достоверными и непротиворечивыми («шум» отсутствует) или в выборке возможны ошибки (наличие «шума»). Если в таблице обучения имеется «зашумленный» объект, то стратегия избежания переподгонки должна быть направлена на «отбрасывание» такого объекта: в идеальном случае именно на нем при ошибке обучения должна остановиться подгонка.

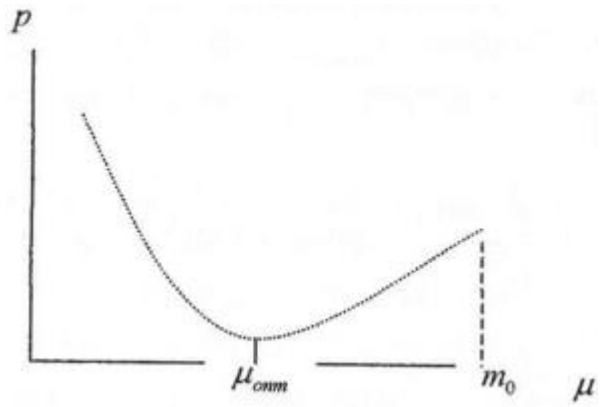

Рис. 1. Кривая в координатах «ошибка» — «сложность РД»

Таким образом, любая стратегия, направленная на избежание переподгонки представляет собой одну из форм смещения и, по существу, должна ухудшать качество распознавания [12]. В защиту стратегий редукции выступают многочисленные эксперименты. Однако эффективность стратегий редукции можно объяснить не столько природой эвристических стратегий, сколько удачным подбором областей для проведения экспериментов, а также подстройкой стратегий редукции под эмпирические данные за счет варьируемого параметра — уровня значимости. Следует отметить, что нет четкого теоретического обоснования, как выбирать значения для уровня значимости, в связи, с чем классификатор типа РД оказывается либо недостаточно, либо чрезмерно редуцированным.

Отсутствие достаточного теоретического обоснования эффективности стратегий редукции, а также выявление их недостатков в ходе проведения многочисленных экспериментов привели к поиску других стратегий избежания переподгонки [7, 10].

Представляется целесообразным проведение дальнейших исследований, напрвленных на строгое теоретическое обоснование редукции и стратегий подгонки РД.

Список литературы

1. **Breslow L.A., Aha D.W.** Comparing Tree-Simplification Procedures// Navy Center for Applied Research in Artificial Intelligence, Technical Report No. AIC-96-015, pp. 1-10, 1996.
2. **Breslow L.A., Aha D.W.** Simplifying Decision Trees: A Survey// Knowledge Engineering Review 12, pp. 1-40, 1997.
3. **Dong M., Kothari R.** Classifiability Based Pruning of Decision Trees// Proc. International Joint Conference on Neural Networks (IJCNN), Vol. 3, pp. 1739-1743, 2001.
4. **Elomaa T., Kaariainen.** An Analysis of Reduced Error Pruning// Journal of Artificial Intelligence Research 15, pp. 163-187, 2001.
5. **Esposito F., Malerba D., Semeraro G.** A Comparative Analysis of Methods for Pruning Decision Trees// IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5), pp. 476-491, 1997.
6. **Frank E.** Pruning Decision Trees and Lists// Ph. D. thesis, University of Waikato, Department of Computer Science, Hamilton, New Zealand, 2000.
7. **Freund Y., Mansour Y., Schapire R.E.** Why Averaging Can Protect Against Overfitting? // In Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, pp. 1-9, 2001.
8. **Kothari R., Dong M.** Decision Trees for Classification: A Review and Some New Results// Pattern Recognition: From Classical to Modern Approaches, S.R. Pal, and A. Pal (Eds.), Chapter 6, pp. 169-184, World Scientific, 2001.
9. **Malerba D., Esposito F., Semeraro G.** A Further Comparison Methods of Decision Tree Induction// Learning From Data: Artificial Intelligence and Statistics V.D. Fisher and H. Lenz, eds., Lecture Notes in Statistics. Berlin: Springer, no. 112, pp. 365-374, 1996.
10. **Oliver J.J.** On Pruning and Averaging Decision Trees// In Proceedings of the Twelfth International Conference on Machine Learning, pp. 430-437, 1995.
11. **Schaffer C.** When Does Overfitting Decrease Prediction Accuracy in Induced Decision Trees and Rule Sets?// In Proceedings of the European Working Session on Learning (EWSL-91), pp. 192-205, Berlin, 1991.
12. **Schaffer C.** Overfitting Avoidance as Bias// Machine Learning 10, pp. 153-178, 1993.