УДК 004.855.5

Методы снижения взаимозависимости ошибок деревьев решений при обучении лесов решений

И.Л. Кафтанников, А.В. Парасич

Рассматривается задача обучения лесов решений. Приводится анализ существующих методов обучения лесов решений, описываются их преимущества и недостатки. Предлагаются возможные методы снижения взаимозависимости ошибок деревьев.

Ключевые слова: деревья решений, леса решений, машинное обучение, классификация.

Классификация и регрессия на основе деревьев решений используется в задачах распознавания образов, информационного поиска, прогнозировании временных рядов, классификации текстов, распознавания речи и др. При этом от качества обучения деревьев существенно зависит правильность решения задачи и практическая применимость результатов. Обучение состоит в настройке условий в узлах дерева и ответов в его листах с целью достичь максимального качества классификации. Настройка параметров дерева производится с использованием обучающих данных — множества примеров с известными правильными ответами.

Вместо одиночных деревьев часто используются леса решений — несколько деревьев, результат определяется с помощью голосования (ответом является тот класс, который предсказало наибольшее число деревьев). Это позволяет улучшить результаты классификации.

Пусть заданы конечное множество объектов $X = \{x_1, \cdots, x_L\}$ и алгоритмов $A = \{a_1, \cdots, a_D\}$ и бинарная функция потерь $I: A \times X \to \{0, 1\}$, $I(a, x) = 1$ тогда и только тогда, когда алгоритм допускает ошибку на объекте x. Число ошибок алгоритма a на выборке X определяется как $n(a, X) = \sum_{x \in X} I(a, x)$. Частота ошибок алгоритма на выборке определяется как $v(a, X) = n(a, x)/|X|$. Под качеством классификации понимается частота ошибок алгоритма на контрольной выборке [1].

Для того чтобы лес решений повышал качество, необходимо обеспечить независимость ошибок деревьев (иначе говоря, если все деревья будут ошибаться на одних и тех же примерах, не будет выигрыша от использования леса). Верхняя граница ошибки обобщения (вероятности неправильной классификации) для леса определяется как

$$GE = p\frac{1 - s^2}{s^2},$$

где GE – ошибка обобщения, s – качество классификации дерева, p – средняя попарная корреляция между деревьями [2].

Можно выделить следующие причины ошибок деревьев:

- случайные ошибки при измерении признака — редкие незакономерные ошибки при вычислении отдельных признаков; использование лесов решений повышает устойчивость к данному виду ошибок;

- систематические ошибки при измерении признака — ошибки при вычислении некоторых признаков, происходящие в определенных условиях; леса решений снижают ущерб от недостаточно надёжных признаков; правильно выбранный метод обучения леса может помочь в устранении ошибок данного вида;

- недостаточность обучающих данных — в результате выбираются неверные решающие правила, или несколько разных решающих правил одинаково подходят на учебных данных; при использовании лесов решений повышается вероятность того, что зависимость будет восстановлена верно;

- сложная форма разделяющей поверхности классов — закономерности в исходных данных слишком сложны, требуются деревья большой глубины, что может приводить к переобучению; требуется больший объем данных для правильного обучения; устранение ошибок данного вида сильнее всего зависит от метода обучения.

Если обучать деревья на одном и том же обучающем множестве одним и тем же методом, получатся одинаковые или очень похожие деревья (если в обучении используется элемент случайности). Поэтому требуется принимать специальные меры для достижения непохожести деревьев. Способы повышения независимости ошибок деревьев можно разделить на два вида — методы на основе предположений и методы, явно максимизирующие качество классификации. Методы первого типа основаны на некоторых базовых предположениях относительно процесса обучения. Методы второго типа обучают несколько вариантов модели и выбирают ту, которая дает наилучшее качество классификации. Общий недостаток методов данного типа — большое время обучения.

Методы на основе предположений

Метод Bagging [3] — каждое дерево обучается на собственном подмножестве обучающей выборки, выбранном случайно. Недостаток данного метода – при росте размера обучающей выборки эффект пропадает, так как подвыборки становятся все более похожими (поскольку взяты из одного вероятностного распределения, а влияние случайных — ослабевает).

Метод случайных подпространств — каждое дерево выбирает признаки в вершинах дерева из подмножества признаков, не пересекающегося с подмножествами признаков других деревьев. Данный метод позволяет хорошо справляться с систематическими ошибками признаков, особенно эффективен, когда признаки в разных деревьях имеют разную природу

(например, видео и звуковые признаки). Недостаток — значения признаков часто взаимозависимы.

Метод Boosting [4] — тренировочным примерам назначаются веса в зависимости от их сложности. Вначале все веса равны единице. Обучается одно дерево, с его помощью производится классификация тренировочных примеров. Веса примеров, классифицированных правильно — снижаются, классифицированных неправильно — повышаются. Следующее дерево леса строится с учетом обновленных весов, и так далее до достижения требуемой ошибки классификации или количества деревьев. Метод позволяет значительно улучшать результаты работы отдельных деревьев. Недостаток – требуется много деревьев для достижения приемлемого качества классификации, что не позволяет использовать данный метод для обучения лесов, работающих в условиях реального времени. Потребность в большом количестве деревьев возникает из-за того, что требуется выбирать, насколько изменять вес примеров, из априорных соображений. При этом для разных примеров оптимальным будет разное изменение веса, и не всегда можно определить разумное изменение веса без предварительного обучения.

Методы, явно максимизирующие качество классификации

Генетический алгоритм – имеется популяция деревьев, на очередном шаге происходит случайная мутация некоторых деревьев и отбор тех деревьев, которые дают наименьшую ошибку классификации при использовании их в качестве деревьев леса. Недостаток – слишком большое время обучения.

ComBoost [5] — после обучения очередного дерева для каждого учебного примера вычисляется отступ — степень уверенности леса в классификации данного примера [6]. Объекты со слишком большим отступом считаются выбросами и удаляются из выборки. Далее перебирается нижняя граница отступа. Примеры с отступом ниже минимального считаются простыми для классификации, и не участвуют в обучении нового дерева. После нескольких итераций выбирается дерево, которое больше всего снижает ошибку классификации леса.

Предлагаемые методы

Методы, явно максимизирующие качество классификации, долго работают на обучающих выборках большого размера. Если размер выборки, необходимый для достижения достаточного качества классификации, таков, что обучение одного варианта модели занимает несколько дней или недель, то обучение много вариантов моделей затруднено. Однако общие закономерности в данных можно установить, используя подвыборку небольшого размера. При обучении деревьев решений можно сгенерировать набор вершин дерева (под вершиной здесь понимается первые несколько

уровней дерева), и обучить дерево на подвыборке небольшого размера. Затем отобрать наилучшие деревья и использовать их вершины в качестве вершин деревьев итогового леса, обученного по всей выборке.

Качество дерева можно определять по методу скользящего контроля. Исходное множество примеров с известными правильными ответами разбивается на две подвыборки: обучающую и контрольную. После настройки параметров дерева по обучающей выборке вычисляется средняя ошибка классификации на контрольной выборке, которая и используется в роли меры качества дерева. Стоит отметить, что такого вида оценка качества классификаторов обладает рядом принципиальных недостатков, так как обучающая и контрольная выборка сгенерированы из одного и того же вероятностного распределения, которое не всегда точно отражает истинное распределение, кроме того, оценка зависит от баланса примеров разного вида в выборке. Подобная методика тестирования не позволяет обнаружить ошибки в формировании обучающей выборки, которые часто оказываются более критичны, чем недостатки алгоритма обучения. Поэтому лучше оценивать качество классификации по выборке из независимого источника, или как среднее по нескольким выборкам из разных независимых источников, так как каждая выборка имеет свои особенности формирования.

Таким образом, в вершинах разных деревьев гарантированно окажутся разные условия. В результате применения данного метода (назовём его «Метод сокращённого обучения вершин») качество улучшается. Однако в ходе построения дерева все равно могут возникать вершины с похожими множествами обучающих примеров, в которых будут выбраны одинаковые условия, что является недостатком метода.

При этом для определения оптимального размера вершины дерева всё равно потребуется проводить обучение несколько раз. Однако, в отличие от методов, явно максимизирующих качество классификации, выбор размера вершины не является основой алгоритма и может быть произведён однократно или задан из априорных соображений.

Чтобы преодолеть данный недостаток, можно явно поощрять обучение непохожих друг на друга деревьев (назовём данный метод «Decorrelation Boosting»). При выборе условия в вершину дерева используется следующий критерий качества разбиения учебных примеров на два множества:

$$I = \frac{|L|}{|L| + |R|} * H(L) + \frac{|R|}{|L| + |R|} * H(R),$$

где $|L|$ — мощность первого подмножества, $|R|$ — мощность второго подмножества, $H(L)$, $H(R)$ — информативность подмножества, которая может выражаться энтропией Шеннона или индексом Гини.

Можно добавить штраф за генерацию подмножеств обучающих примеров, похожих на подмножества, ранее полученные при обучении предыдущих деревьев леса:

$$H\acute{(x)} = H(x) + T_c(x),$$

где $T_c(x)$ — штраф за похожесть подмножества x на ранее встречавшиеся:

$$T_C(x) = \max_{i \in P} C_{i,x}) * (ERR(i) + \varepsilon),$$

где P — подмножества примеров, полученные в результате обучения предыдущих деревьев, x — текущее подмножество, *ERR(i)* — ошибка классификации в вершине i на обучающем множестве, ε — процент случайной ошибки дерева;

$$\mathrm{C}_{i,\mathrm{x}} = \frac{\min(CP_{i,x}, CP_{x,i})}{|CP_{i,x} - CP_{x,i}|+1},$$

где $CP_{i,x}$ — доля элементов множества i, содержащихся в множестве x.

Если для каждой пары обучающих примеров известна их похожесть согласно некоторой метрике, можно использовать данную информацию при определении похожести двух подмножеств.

Еще один возможный подход — использовать методы, явно максимизирующие качество классификации для обучения вершин с малым числом обучающих примеров. Обычно большой процент ошибок возникает именно в вершинах с малым объемом обучающего множества из-за недостаточности данных. В то же время, обучение данных вершин занимает меньше всего времени.

Выводы

Рассмотрены причины, приводящие к ошибкам деревьев решений. Проведен подробный анализ существующих методов снижения взаимозависимости ошибок деревьев решений при обучении леса решений, выявлены их преимущества и недостатки. Предложена классификация методов снижения взаимозависимости ошибок деревьев решений. Предложены новые методы решения данной проблемы. Полученные результаты легко обобщаются на случай композиций произвольных классификаторов.

Предложенные способы решения проблемы взаимозависимости ошибок классификаторов позволят одновременно повысить качество классификации и объединить достоинства существующих методов.

Библиографический список

1. Воронцов К.В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы. / К.В. Воронцов // Математические методы распознавания образов — 2011 — С. 40–43.

2. Breiman L. Random Forests. / L. Breiman // Machine Learning — 2001 — Vol. 45(1), P. 5–32.

3. Breiman L. Bagging Predictors. / L. Breiman // Machine Learning — 1996 — Vol. 24, No. 2. P. 123–140.

4. Freund Y. Experiments with a New Boosting Algorithm. / Y. Freund, R. E. Schapire // International Conference on Machine Learning — 1996 — P. 148–156.

5. Маценов А.А. Комитетный бустинг: минимизация числа базовых алгоритмов при простом голосовании. / А.А. Маценов // Математические методы распознавания образов — 2013 — С. 180–183.

6. Mason L. Direct Optimization of Margins Improves Generalization in Combined Classifiers. / L. Mason, P. Bartlett, J. Baxter // Proc. of the 1998 conf. on Advances in Neural Information Processing Systems II — 1999 — P. 288–294.

Methods of decrease in interdependence of errors of decisions trees when training the decisions woods

I.L. Kaftannikov, A.V. Parasich

The problem of training of the decisions woods is considered. The analysis of the existing methods of training of the decisions woods is provided, their advantages and shortcomings are described. Possible methods of decrease in interdependence of errors of trees are offered.

Keywords: decisions trees, decisions woods, machine training, classification.

Библиографический список

1. Vorontsov K.V. [Combinatory Theory of Retraining: Results, Appendices and Open Problems]. *Mathematical methods of recognition of images*, 2011, pp. 40–43. (in Russ.)

2. Breiman L. Random Forests. *Machine Learning*, 2001, Vol. 45(1), pp. 5–32.

3. Breiman L. Bagging Predictors. *Machine Learning*, 1996, Vol. 24, No. 2, pp. 123–140.

4. Y. Freund, R. E. Schapire Experiments with a New Boosting Algorithm. *International Conference on Machine Learning*, 1996, pp. 148–156.

5. Matsenov A.A. [Committee Busting: Minimization of Number of Basic Algorithms at Simple Vote]. *Mathematical methods of recognition of images*, 2013, pp. 180–183. (in Russ.)

6. L. Mason, P. Bartlett, J. Baxter Direct Optimization of Margins Improves Generalization in Combined Classifiers. *Proc. of the 1998 conf. on Advances in Neural Information Processing Systems II*, 1999, pp. 288–294.