УДК 519.7

ПРИМЕНЕНИЕ ЭМПИРИЧЕСКОГО РЕШАЮЩЕГО ЛЕСА ДЛЯ ФИЛЬТРАЦИИ ОБУЧАЮЩИХ ДАННЫХ

Дюличева Ю.Ю.

Таврический национальный университет им. В.И.ВЕРНАДСКОГО,
кафедра информатики
просп.ВЕРНАДСКОГО, 4, г.Симферополь, Украина, 95007
e-mail: *DYULICHEVA@MAIL.RU*

Аннотация

A clear outliers detection algorithm (training sample filtering algorithm) based on the empirical decision forest with branches' *rank r* is proposed in the paper. The generalization performance increase of decision tree constructed after filtering in comparison with the decision tree, constructed before filtering, is grounded empirically.

Введение

Серьёзной проблемой, характерной для большинства алгоритмов обучения, является *переподгонка* (overfitting) или *переобучение*. Эффект переподгонки возникает тогда, когда алгоритм все точнее и точнее «настраивается» на обучающую выборку, но при этом все хуже и хуже распознает объекты, не участвовавшие в обучении, т.е. минимизация эмпирического риска вплоть до нуля приводит к ухудшению обобщающей способности алгоритма вне обучающей выборки. По сути, переподгонка противоречит принципу эмпирической индукции — чем больше произведено наблюдений, тем точнее будет прогноз. Такое противоречие принято называть *парадоксом индуктивного обобщения* или *парадоксом переобучения*. В худшем случае, переподгонка может привести к потере способности алгоритма обучения отображать «структуру» обучающей выборки в виде набора закономерностей и обобщать обучающие данные.

Эффект переподгонки — это проблема, которая свойственна не только сложным алгоритмам обучения из теории обучаемых систем, но и хорошо известна в статистике.

Эффект переобучения можно увидеть, наблюдая за «поведением» обучающего алгоритма на контрольной выборке — число ошибок на контроле сначала уменьшается, а затем начинает резко увеличиваться.

Волкер Наннен (Volker Nannen) приводит любопытную демонстрацию эффекта переобучения [1]. На рисунке 1 изображены две кривые: первая — «отрезок» аттрактора Лоренца (сейчас существенно только то, что аттрактор Лоренца не является полиномом), вторая кривая — «оптимальный» полином 43 - ей степени (изображен в виде непрерывной линии с меньшими колебаниями, чем у аттрактора Лоренца). Известно, что аттрактор Лоренца может быть аппрок-симирован полиномом. Кривые

построены по обучающей выборке, содержащей 300 объектов, а контроль проводился на 3000 объектов. И обучающая, и контрольная выборки независимы и одинаково распределены. Распределение по оси x равномерное на всем отрезке $[0; 10]$.

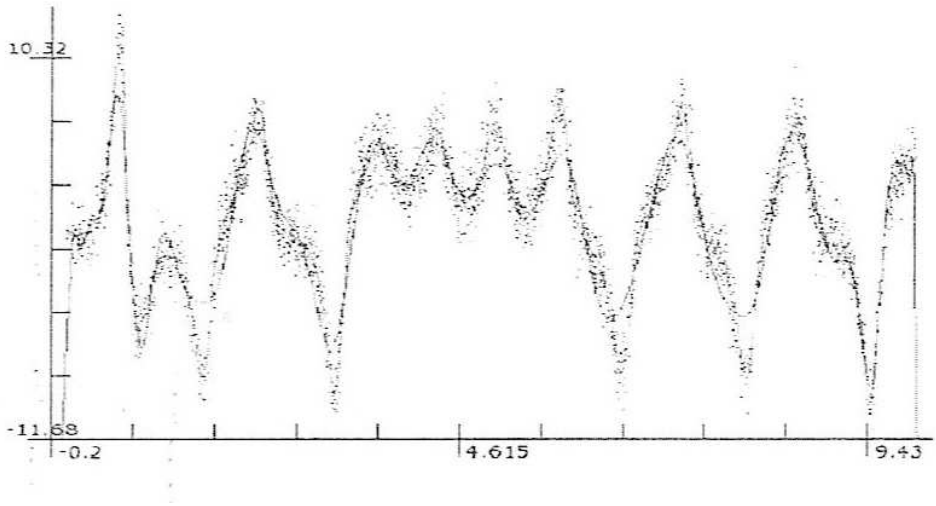


Рис. 1. Расположение точек, соответствующих 300 объектам из обучающей выборки

.

Исследуем обобщающую способность полиномов степени от 0 (константа) до 60, рассмотрев величину их средней квадратической ошибки на контрольном множестве из 3000 объектов. На рисунке 2 на оси x выписаны степени полиномов; ось y показывает ошибку обобщения (ошибку на контроле) и имеет логарифмический масштаб. Слева на графике (рис. 2) первое значение соответствует полиному нулевой степени (константе). Его средняя квадратическая ошибка на контрольной выборке из 3000 объектов равна $18(\sigma^2 = 18)$. Справа от него наблюдается медленное уменьшение величины ошибки на контроле до тех пор, пока она не достигает глобального минимума для полинома 43-ей степени при $\sigma^2 = 2,7$. После этого наблюдается то резкое уменьшение, то резкое увеличение величины ошибки на контроле до локального максимума, который начинает превышать даже начальную величину ошибки $\sigma^2 = 18$. Таким образом, начиная с полинома 44-ой степени наблюдается эффект переобучения, несмотря на то, что сначала (вплоть до полинома 43-ей степени) результат был вполне ожидаемым — величина ошибки на контроле уменьшалась с ростом числа параметров модели. Чем больше число параметров выбранной модели обучения (число параметров возрастает с ростом степени полинома), тем «точнее» кривые, соответствующие полиномам, .

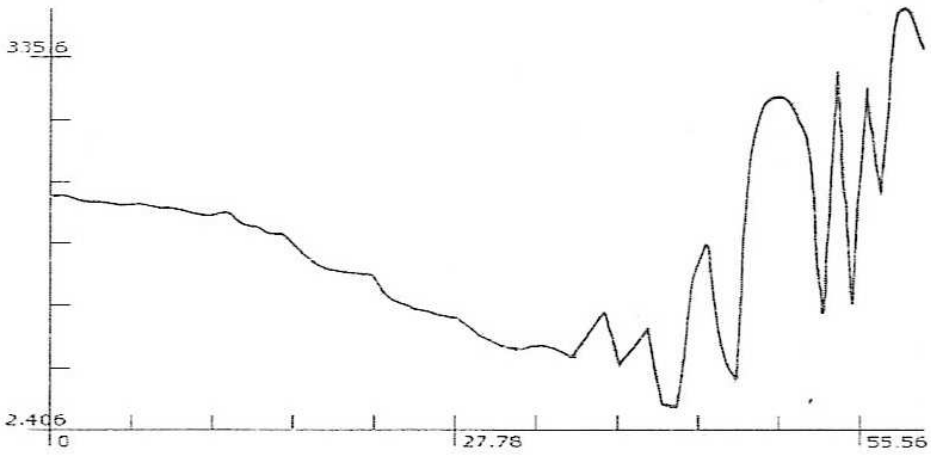


Рис. 2. Демонстрация эффекта переобучения на 3000 контрольных объектов

«проходят» через обучающие объекты, изображенные точками на рисунке 1. Однако выполнение более точной «настройки», чем обеспечивает полином 43 - ей степени, не требуется.

Наиболее распространенным методом предотвращения переобучения (overfitting avoidance method) для решающих деревьев является метод редукции (pruning) или отсечения «избыточных» ветвей. В работе [2] была предложена индуктивная модель — эмпирический решающий лес —, позволяющая предотвращать переобучение эффективнее, чем одно решающее дерево (РД).

Целью данной работы является разработка алгоритма фильтрации обучающих данных. Фильтрация обучающей выборки — это еще один достаточно эффективный метод предотвращения переобучения на обучающей выборке — алгоритм-фильтр осуществляет поиск и удаление «зашумленны» объектов - «выбросов» из обучающей выборки, что способствует повышению обобщающей способности алгоритма обучения, построенного на редуцированной обучающей выборке, полученной из исходной путем удаления «выбросов», В данной работе для выявления «выбросов» в обучающей выборке предлагается использовать эмпирический решающий лес (ЭРЛ) ранга r.

В первом разделе предлагается модификация алгоритма синтеза эмпирического решающего леса, направленная на удаление выбросов из обучающей выборки. Во втором разделе приводятся результаты эмпирического сравнения обобщающей способности решающего дерева до и после процесса фильтрации.
В заключении подводятся итоги работы и указываются направления дальнейших исследований.

1. Эвристический алгоритм фильтрации обучающих данных на основе эмпиреческого решающего леса ранга r

Пусть X^l — обучающая выборка, $y^* : X \to Y$ — неизвестная целевая зависимость, X — множество всех допустимых объектов (генеральная совокупность), Y — множество ответов (меток классов). Известны значения целевой зависимости на обучающей выборке, т.е. $y_i = y^*(x_i)$ для любого $i = 1,2,\ldots,l$.

Требуется построить алгоритм a : $X \to$ У, восстанавливающий неизвестную целевую зависимость y^*. В качестве *алгоритма a* в данной работе рассматривается решающее дерево, а эмпирический решающий лес используется для выявления чистых выбросов в обучающей выборке X^l . После удаления из обучающей выборки чистых выбросов снова строится решающее дерево.

Определение 1. Объект $\tilde{x} \in X^l$ называется чистым выбросом относительно модели A алгоритмов обучения, если $L(\mathrm{a}(\tilde{x}), y^*(\tilde{x})) = l$ для любого алгоритма a $\in$ A, где $L(\mathrm{a}(\tilde{x}), \mathrm{y}^*(te\tilde{x}tit\mathrm{x}) = [\mathrm{a}(\tilde{x}) \neq y^*(\tilde{x})]$ - функция потерь от ошибки при классификации объекта $\tilde{x} \in X^l$ алгоритмом a $\in$ A.

Напомним, что областью некомпетентности или областью отказа решающего дерева называется интервал, соответствующий ветви, ранг которой превышает некоторое пороговое значение. Предполагается, что в область отказа решающего дерева попадают объекты, вызывающие излишнее усложнение структуры решающего дерева и образующие множество чистых выбросов относительно эмпирического решающего леса ранга r.

Исходя из этого предположения, используем алгоритм синтеза эмпирического решающего леса ранга r, введенный в работе [2], для выявления чистых выбросов в обучающей выборке.

Поясним некоторые процедуры, используемые в алгоритме 1.

$BuildTree(d_i)$ — процедура построения i-го решающего дерева леса. В работе [3] отмечалось, что дерево d_i строится, в первую очередь, на объектах из пересечения множеств выбросов предыдущих $i-1$ деревьев ($i \geq 1$) и признаках, которые еще не участвовали в обучении (при условии, что указанные множества не пусты), а затем решающее дерево d_i достраивается на объектах, не входящих в пересечение множеств выбросов.

Множество чистых выбросов $FOREST\_OUTLIERS(\mathrm{X}^l, r)$ относительно заданной обучающей выборки и построенного эмпирического решающего леса

АЛГОРИТМ 1. *Модификация алгоритма синтеза эмпирического решающего леса для выявления чистых выбросов* относительно заданной выборки X^l и эмпирического решающего леса ранга r.

Вход:

обучающая выборка X^l, допустимый ранг r;

Выход:

$FOREST\_OUTLIERS(\mathrm{X}^l, r)$ — множество чистых выбросов относительно обучающей выборки X^l и эмпирического решающего леса ранга r;

1: инициализировать счетчик числа деревьев эмпирического решающего леса $i := 1$;
2: **пока** не выполнен критерий останова
3: строить $BuildTree(d_i)$ очередное решающее дерево d_i эмпирического решающего леса;
4: увеличить значение счетчика i на единицу;
5: сформировать множество $TREE\_OUTLIERS(d_i, X^l, r)$ выбросов решающего дерева d_i;
6: сформировать множество чистых выбросов эмпирического решающего леса ранга r:

$$FOREST\_OUTLIERS(\mathrm{X}^l, r) := \cap TREE\_OUTLIERS(d_i, X^l, r)$$

формируется как пересечение множеств выбросов всех деревьев, входящих в состав эмпирического решающего леса. В случае построения корректного эмпирического решающего леса, множество $FOREST\_OUTLIERS(\mathrm{X}^l, r) = \oslash$. Фильтрация обучающей выборки X^l происходит только в случае построения некорректного эмпирического решающего леса и заключается в удалении из выборки X^l всех объектов, входящих во множество $FOREST\_OUTLIERS(X^l, r)$.

Эффективности корректного эмпирического решающего леса ранга r по сравнению с отдельным решающим деревом были посвящены работы [2], [3]. Алгоритм 1 демонстрирует практическую полезность построения некорректного эмпирического решающего леса ранга г для фильтрации обучающих данных.

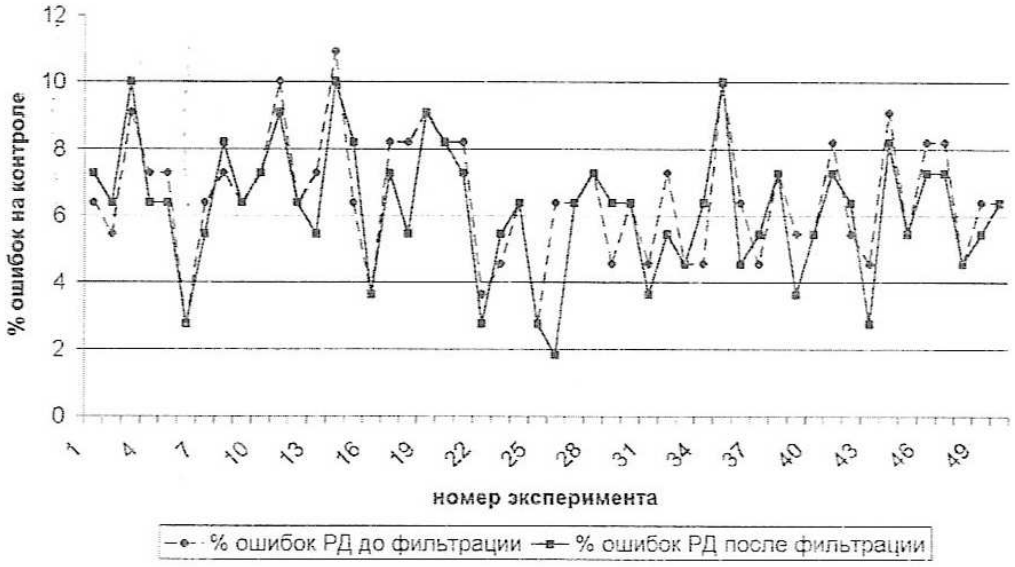
2. Результаты эмпирического сравнения обобщающей способности решающего дерева до и после фильтрации

Для эмпирического обоснования того факта, что решающее дерево после фильтрации дает меньшее число ошибок на контроле, чем решающее дерево, построенное до применения процесса фильтрации на основании эмпирического решающего леса ранга r из базы данных патогенных вибрионов и аэромонад

[2] при проведении каждого эксперимента случайно и независимо выбирались обучающие выборки, состоящие из 255 объектов (70% базы данных). По результатам каждого отдельного эксперимента вычислялся процент ошибок на одной и той же контрольной выборке из 110 объектов (30% базы данных) и для решающего дерева, построенного до фильтрации, и для решающего дерева, построенного после фильтрации.

Рисунки 3 и 4 демонстрируют результаты проведения 50 экспериментов. Изменение процента ошибок на контроле при фильтрации обучающей выборки посредством эмпирического решающего леса ранга 5 представлено на рисунке 3. .

Рис. 3. % ошибок РД до и после фильтрации посредством ЭРЛ ранга 5

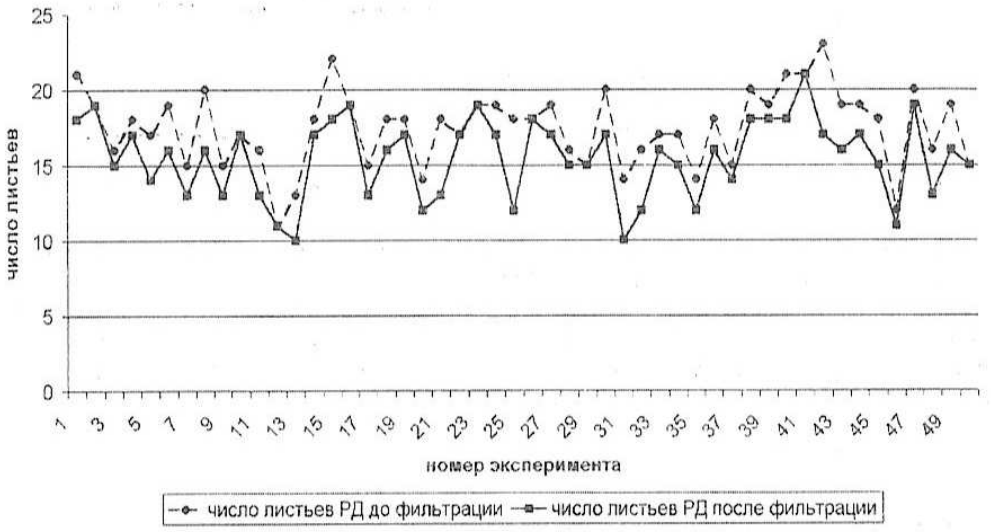


Из рисунка 3 видно, что в 39 случаях из 50 решающее дерево после фильтрации имеет процент ошибок на контроле не больше, чем решающее дерево до фильтрации, при этом структура (число листьев) решающего дерева после фильтрации значительно упрощается.

Изменение числа листьев у решающих деревьев, построенных до и после фильтрации с помощью эмпирического решающего леса ранга. 5, представлено на рисунке 4.

.

Рис. 4. Число листьев РД до и после фильтрации посредством ЭРЛ ранга 5



Заключение

В работе предложен эвристический алгоритм фильтрации обучающей выборки, основанный на построении некорректного эмпирического решающего леса ранга r и направленный на предотвращение переобучения. В дальнейшем предполагается разработка теоретически обоснованного критерия фильтрации обучающих данных на основе метрических свойств булевых функций.

Автор благодарит профессора В. И. Донского за внимание к работе.

Список литературы

1. Volker Nannen The Paradox of Overfitting master's thesis, s1030841, Faculty of Artificial Intelligence in Groingen. — 2003. — 78p.
2. Дюличева Ю.Ю. О программной реализации и апробации и алгоритма *DFBSA* синтеза эмпирического решающего леса. Таврический вестник информатики и математики. — 2003. — № 2. — С. 35—43.
3. Дюличева Ю.Ю. Оценка *VCD* r -редуцированного эмпирического леса. Таврический вестник информатики и информатики. — 2003. — № 1. — С.31—42.
4. Brockley C.E., Friedl M.A. Identifying Mislabeled Training Data. Journal of Artificial Intelligence Research. — 1999. — no.11. — P.131—167.