**Прунинг деревьев принятия решений на основе принципа сокращенной ошибки (REP)**

На протяжении десятилетий усилия исследователей в области обучения распознаванию на основе решающих деревьев (РД) были направлены на разработку стратегий избегания 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. 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 избегания переподгонки называется редукцией

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 уменьшение числа листьев. Процесс редукции направлен на сокращение зависимости РД от «зашумленных» данных.

Предредукция и постредукция — 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 Х2 или точный критерий Фишера). Предредукция не является эффективным методом избегания 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.

Согласно стратегии *RЕР* обучающее множество разбивается на два подмножества: множество синтеза РД (70% от обучающего множества) и множество редукции РД (30% от обучающего множества). РД *Т* 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(Т(t))*, допускаемых поддеревом с корневой вершиной # 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)) > I(t*), то поддерево с корневой вершиной t редуцируют. Далее процесс редукции применяется к полученному редуцированному дереву до тех пор, пока для всех внутренних 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 RЕР 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 RЕР 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 RЕР. Пустые поддеревья связаны с исследованием интервалов наименьшей размерности, соответствующих ветви дерева и покрывающих сравнительное небольшое число объектов обучающего множества.

Пусть *rL* - ошибка листа с меткой мажоритарного класса, который заменит в процессе редукции поддерево *T(t)* с корнем в вершине *t,* тогда *rL* зависит только от (распределения классов) объектов, которые достигают корневой вершины *t* поддерева *T(t)*. Другими словами, структура дерева, расположенная выше поддерева Т(Е) определяем величину ошибки *rL*. Пусть *rT(t)* - ошибка поддерева *T(t)* после проведения редукции в этом поддереве согласно стратегии RЕР, тогда величина ошибки *rT(t)* либо уменьшилась либо осталась неизменной, Таким образом, *rL* и *rT(t)* -

независимые случайные величины. Следовательно, вероятность события *rT(t)* < *rL*

увеличивается в процессе редукции дерева *T(t)*. Это смещение в распространении ошибки свойственно стратегии RЕР.

В результате реализации стратегий RЕР будет получено редуцированное РД с наибольшей точностью классификации на объектах множества редукции. RЕР 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.