ЛАБОРАТОРНАЯ РАБОТА 12
КЛАССИФИКАЦИЯ С ПОМОЩЬЮ ДЕРЕВЬЕВ РЕШЕНИЙ

Деревья решений применяются для решения задачи классификации. Дерево представляет собой иерархический набор условий (правил), согласно которым данные относятся к тому или иному классу. В построенном дереве присутствует информация о достоверности того или иного правила. Рассчитывается значимость каждого входного поля.

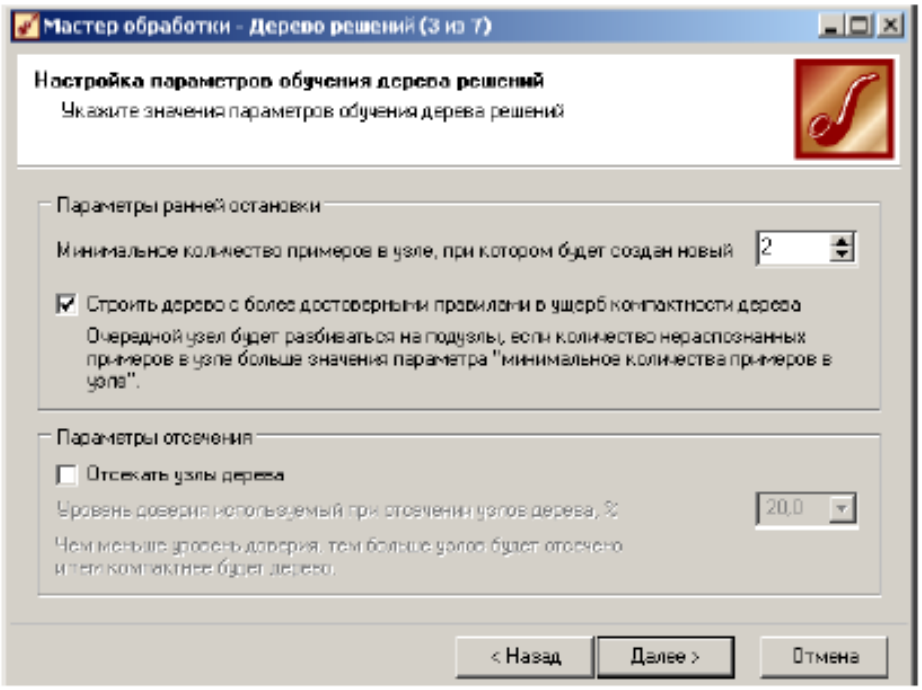
Исходные данные

Пусть аналитик имеет данные по тому, как голосуют депутаты конгресса США по различным законопроектам. Также известна партийная принадлежность каждого депутата – республиканец или демократ. Перед аналитиком поставлена задача: классифицировать депутатов на демократов и республиканцев в зависимости от того, как они голосуют. Данные по голосованию находятся в файле "Vote.txt". Таблица содержит следующие поля : "Код" – порядковый номер, "Класс" – класс голосующего (демократ или республиканец), остальные поля информируют о том, как голосовали депутаты за принятие различных законопроектов ("да" , "нет" , "воздержался").

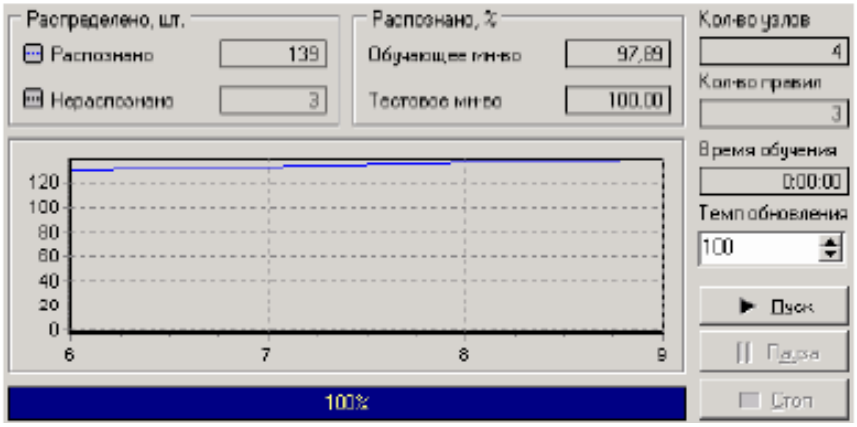
Классификация на демократов и республиканцев

Для решения задачи запустим Мастер обработки. Выберем в качестве обработки дерево решений. В Мастере построения дерева решения на втором шаге настроим поле "Код" информационным, "Класс" выходным, остальные поля входными.

Далее предлагается настроить способ разбиения исходного множества данных на обучающее и тестовое. Зададим случайный способ разбиения, когда данные для тестового и обучающего множества берутся из исходного набора случайным образом. На следующем шаге Мастера предлагается настроить параметры процесса обучения, а именно минимальное количество примеров, при котором будет создан новый узел (пусть узел создается, если в него попали два и более примеров), а также предлагается возможность строить дерево с более достоверными правилами. Включим данные опции.



На следующем шаге Мастера выбираем процесс построения дерева решения в автоматическом режиме или интерактивном (полуавтоматическом). Выберем сначала автоматический режим построения и запустим. Далее можно увидеть информацию о количестве распознанных примеров.



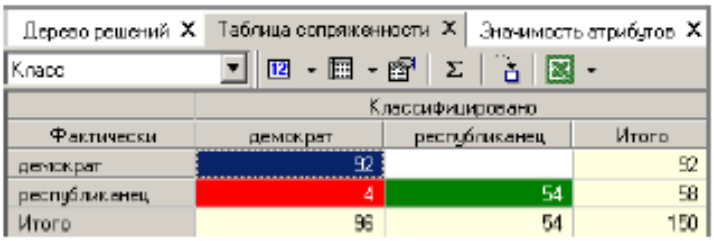
После построения дерева можно увидеть, что почти все примеры и на обучающей и на тестовой выборке распознаны.

Перейдем на следующий шаг Мастера для выбора способа визуализации полученных результатов. Основной целью аналитика является отнесение депутата к той или иной партии. Механизм отнесения должен быть таким, чтобы депутат указал, как он будет голосовать за различные законопроекты, а дерево решений ответит на вопрос, кто он – демократ или республиканец. Такой механизм предлагает визуализатор "Что-если". Не менее важным является и просмотр самого дерева решений, на котором можно определить, какие факторы являются более важными (верхние узлы дерева), какие второстепенными, а какие вообще не оказывают влияния (входные факторы,

вообще не присутствующие в дереве решений). Поэтому выберем также и визуализатор "Дерево решений". Формализованные правила классификации, выраженные в форме "Если <Условие>, тогда <Класс>", можно увидеть, выбрав визуализатор "Правила (дерево решений)". Часто аналитику бывает полезно узнать, сколько примеров было распознано неверно, какие именно примеры были отнесены к какому классу ошибочно. На этот вопрос дает ответ визуализатор "Таблица сопряженности". Очень важно знать, каким образом каждый фактор влияет на классификацию. Такую информацию предоставляет визуализатор "Значимость атрибутов".

Результат

Проанализируем данные при помощи имеющихся визуализаторов. Для начала посмотрим на таблицу сопряженности.

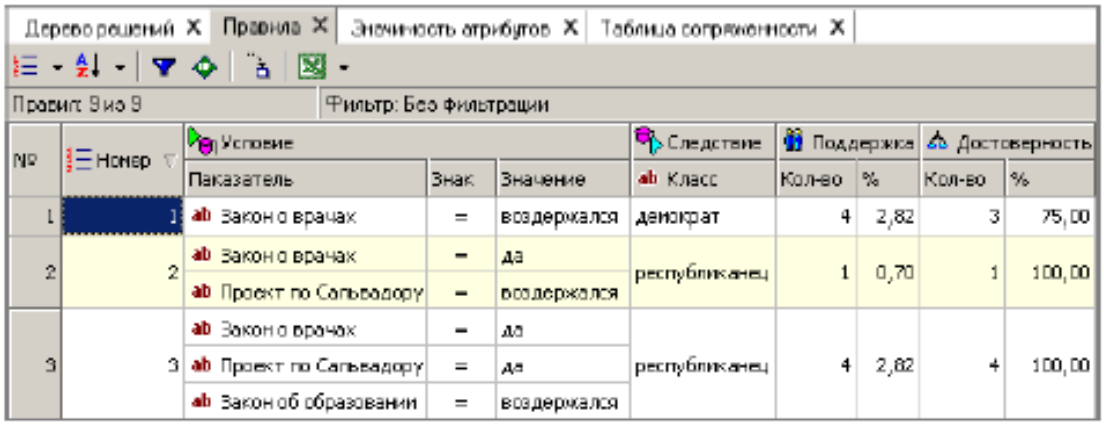

Дерево решений | Таблица сопряженности | Значимость атрибутов

Класс

Фактически	Классифицировано: демократ	республиканец	Итого
демократ	92		92
республиканец	4	54	58
Итого	96	54	150

По диагонали таблицы расположены примеры, которые были правильно распознаны, в остальных ячейках - те, которые были отнесены к другому классу. В данном случае дерево правильно классифицировало практически все примеры.

Перейдем к основному визуализатору для данного алгоритма – "Дерево решений". Как видно, дерево решений получилось не очень громоздкое, большая часть факторов (законопроектов) была отсечена, т.е. влияние их на принадлежность к партии минимальна или его вообще нет (по-видимому, по этим вопросам у партий нет принципиального противостояния).


Дерево решений | Правила | Значимость атрибутов | Таблица сопряженности

Правил: 3 из 3 | Фильтр: Без фильтрации

№	Номер	Условие: Показатель	Знак	Значение	Следствие: Класс	Поддержка: Кол-во	%	Достоверность: Кол-во	%
1	1	Закон о врачах	=	воздержался	демократ	4	2,82	3	75,00
2	2	Закон о врачах	=	да	республиканец	1	0,70	1	100,00
		Проект по Сальвадору	=	воздержался					
3	3	Закон о врачах	=	да	республиканец	4	2,82	4	100,00
		Проект по Сальвадору	=	да					
		Закон об образовании	=	воздержался					

Самым значимым фактором оказалась позиция, занимаемая депутатами по пакету законов, касающихся врачей, т. е. если депутат голосует против

законопроекта о врачах, то он демократ (об это можно говорить с полной уверенностью, потому что в узел попало 83 примера). Достоверно судить о том, что депутат – республиканец, можно, если он голосовал за законопроект о врачах, а также за законопроект по Сальвадору, а также был против законопроекта об усыновлении. Данный визуализатор предоставляет возможность просмотра примеров, которые попали в тот или иной узел, а также информацию об узле.

Более удобно посмотреть значимость факторов или атрибутов в визуализаторе "Значимость атрибутов".

Целевой атрибут: Класс

№	Атрибут	Значимость, %
4	Закон о врачах	92.207
16	Проект по экспорту	3.498
5	Проект по Сальвадору	2.455
3	Проект по усыновлению	1.840
12	Закон об образовании	0.000
11	Проект по альтернативным источникам топлива	0.000
13	Проект по фондам	0.000
15	Проект по таможенным пошлинам	0.000
14	Проект по преступности	0.000
10	Закон об иммигрантах	0.000
6	Закон о религиях	0.000
2	Проект по водным ресурсам	0.000
1	Проект по инвалидам	0.000
9	Проект по ракетам	0.000
8	Проект помощи Никарагуа	0.000
7	Антиспутниковый проект	0.000

С помощью данного визуализатора можно определить, насколько сильно выходное поле зависит от каждого из входных факторов. Чем больше значимость атрибута, тем больший вклад он вносит при классификации. В данном случае самый большой вклад вносит закон о врачах, как и было сказано выше.

На визуализаторе "Правила" представлен список всех правил, согласно которым можно отнести депутата к той или иной партии. Правила можно сортировать по поддержке, достоверности, фильтровать по выходному классу (к примеру, показать только те правила, согласно которым депутат является демократом с сортировкой по поддержке).

Дерево решений | Правила | Значимость атрибутов | Таблица сопряженности

Правил: 9 из 9 | Фильтр: Без фильтрации

№	Номер	Условие			Следствие	Поддержка		Достоверность	
		Показатель	Знак	Значение	Класс	Кол-во	%	Кол-во	%
1	1	Закон о врачах	=	воздержался	демократ	4	2,82	3	75,00
2	2	Закон о врачах	=	да	республиканец	1	0,70	1	100,00
		Проект по Сальвадору	=	воздержался					
3	3	Закон о врачах	=	да	республиканец	4	2,82	4	100,00
		Проект по Сальвадору	=	да					
		Закон об образовании	=	воздержался					

Данные представлены в виде таблицы. Полями этой таблицы являются:

Данные представлены в виде таблицы. Полями этой таблицы являются:

– номер правила,

– условие, которое однозначно определяет принадлежность к партии,

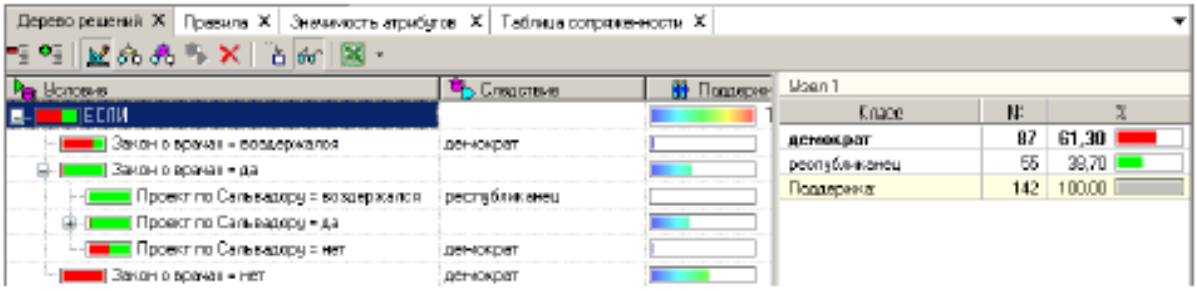
– следствие – то, кем является депутат, голосовавший согласно этому условию,

– поддержка – количество и процент примеров из исходной выборки, которые отвечают этому условию,

– достоверность – процентное отношение количества верно распознанных примеров, отвечающих данному условию, к общему количеству примеров, отвечающих данному условию.

Исходя из данных этой таблицы, аналитик может сказать, что именно влияет на то, что депутат является демократом или республиканцем, какова цена этого влияния (поддержка) и какова достоверность правила. В данном случае совершенно очевидно, что из всего списка правил с достаточно большим доверием можно отнестись к двум : правилу №9 и правилу №7. Таким образом, получается, что демократы принципиально против законопроектов, касающихся врачей. Республиканцы же, наоборот, за принятие этих законопроектов и также за принятие законопроекта по Сальвадору, но категорически против законопроектов по усыновлению.

Теперь аналитик может точно сказать, кто есть кто.



Но иногда аналитик считает правильным построить дерево решений исходя из своих соображений или внести некоторую корректировку, и тогда необходимо выбрать интерактивный режим построения, в результате чего получим следующее окно дерева решений.

Для внесения изменений в него используют следующие кнопки:

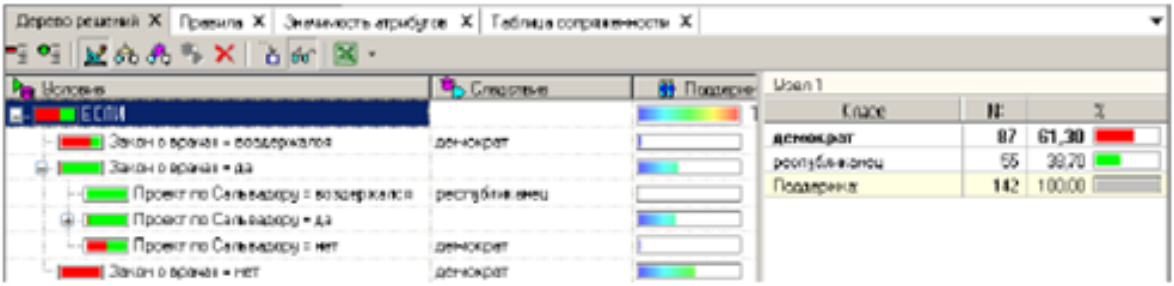
 -включение/выключение интерактивного режима;

 – разбить текущий узел на подузлы;

 – построить дерево решений начиная с текущего узла.

Допустим что аналитик думает, что основное правило которое надо учитывать в построение дерева решений есть проект о ракетах. Тогда для данного построения выберем корневой каталог в дереве решений и нажмем кнопку  и в появившемся окне выберем проект по ракетам. В результате

получим новое дерево решений с новыми правилами и законами.



Выводы

Пример показал простоту и удобство применения деревьев решений для классификации на республиканцев и демократов. Мастер предлагает широкие возможности по настройке процесса построения дерева решений.

Это и настройка назначения столбцов, способов нормализации, настройка источника данных для учителя (тестовое и обучающее множества), настройка количества примеров в узле и настройка достоверности правил. После построения дерева стали видны его достоинства для анализа. Алгоритм сам отсек несущественные факторы, выявил степень влияния тех или иных факторов на результат, описал при помощи формальных правил способ классификации, а также выдал информацию о достоверности и поддержке того или иного правила. Также были продемонстрированы широкие возможности визуализации построенного дерева. Все это говорит о незаменимости деревьев решений для классификации.