ОБУЧЕНИЕ ЧРЕЗ ПОСТРОЯВАНЕ НА ИДЕНТИФИКАЦИОННИ ДЪРВЕТА

Лекция 11

Въведение.

- Ще разгледаме метод, който позволява на компютрите да се обучават чрез формиране на *предварително събрана информация във вид на идентификационно дърво*.
- Най напред ще покажем как да построим едно идентификационно дърво, а след това ще разгледаме как от това идентификационно дърво може да се изградят прегледно множество от правила, построени във вида: една или няколко предпоставки и заключение.

- *Построяването на идентификационно дърво е много широко използван метод за обучение*. Построени са многобройни идентификационни дървета в най-различни приложни области като се започне от медицината (за нуждите например на медицинската диагностика) и се стигне до управление на различни процеси.

- Представете си, че не Ви е известно поради какви причини някои хора изгарят от слънцето след няколко часа прекарани на плажа, а други - се връщат почерняли и щастливи.

Първата стъпка е наблюдението.

Отивате на плажа и започвате да си водите записки за характеристиките на хората, като цвят на косата, височина, тегло, ползване на лосион и разбира се наличие на изгаряния. Целта ви е да използвате наблюдаваните свойства, за да предскажете дали нов човек, дошъл на плажа, ще изгори от слънцето или не.

Нека предположим, че събраната от вас информация е поместена в следната таблица:

Име	Коса	Височина	Тегло	Лосион	Изгаряне
Сара	руса	среден	слаб	не	изгаря
Дана	руса	висок	среден	да	няма
Алекс	кестенява	нисък	среден	да	няма
Анна	руса	нисък	среден	не	изгаря
Емили	червена	среден	тежък	не	изгаря
Пепе	кестенява	висок	тежък	не	няма
Джон	кестенява	среден	тежък	не	няма
Катя	руса	нисък	слаб	да	няма

Вашите наблюдения са включили 4 характеристики(свойства): цвят на косата с три стойности, височина също с три стойности, тегло, което условно сте разделили в три групи и че хората или използват лосион, или не.
Това са 3*3*3*2=54 възможни комбинации.

■Нека на плажа пристигнат Пепа и Пепи.

Пепа е руса, ниска на ръст, със средно тегло, а

Пепи е рус, висок и слаб.

И двамата не използуват лосион.

■Да опитаме да познаем дали ще изгорят на слънцето.

За Пепа намираме ред в таблицата, който съвпада напълно и по четирите показателя (Анна), следователно би могло да се предположи, че тя ще изгори.

Име	Коса	Височина	Тегло	Лосион	Изгаряне
Сара	руса	среден	слаб	не	изгаря
Дана	руса	висок	среден	да	няма
Алекс	кестенява	нисък	среден	да	няма
Анна	руса	нисък	среден	не	изгаря
Емили	червена	среден	тежък	не	изгаря
Пепе	кестенява	висок	тежък	не	няма
Джон	кестенява	среден	тежък	не	няма
Катя	руса	нисък	слаб	да	няма

■Вероятността да попаднем на случая с Пепа в нашата таблица (при равномерно разпределение на вероятностите) е 8/54=0,15, т.е 15%.

Име	Коса	Височина	Тегло	Лосион	Изгаряне
Сара	руса	среден	слаб	не	изгаря
Дана	руса	висок	среден	да	няма
Алекс	кестенява	нисък	среден	да	няма
Анна	руса	нисък	среден	не	изгаря
Емили	червена	среден	тежък	не	изгаря
Пепе	кестенява	висок	тежък	не	няма
Джон	кестенява	среден	тежък	не	няма
Катя	руса	нисък	слаб	да	няма

Пепи е рус, висок и слаб.

За Пепи съвпадението е по два показателя (Дана и Сара). Без да знаем кой показател е най-съществен не бихме могли да направим определено предположение.

■Вероятността характеристиките на новопристигнал посетител на плажа да съвпаднат точно с вече наблюдаваните е много малка, защото в действителност характеристиките са много повече и стойностите им също са много повече.

■ Така че едва ли бихме могли да познаем дали новопристигналия ще изгори или не като търсим съвпадение на характеристиките с някои от вече наблюдаваните от нас, т.е.не е подходящо да търсим класифициране на непознат обект като търсим точно съответствие на характеристиките на този нов обект и характеристиките на вече известните обекти

- Опитите да ползваме някой от известните ни методи за да решим задачата са неуспешни.

- Тъй като не знаем кои свойства са важни, не можем да използваме метода на най-близкия по характеристики съсед, защото може да се получи съвпадение на свойства, които не са сред най-съществени.

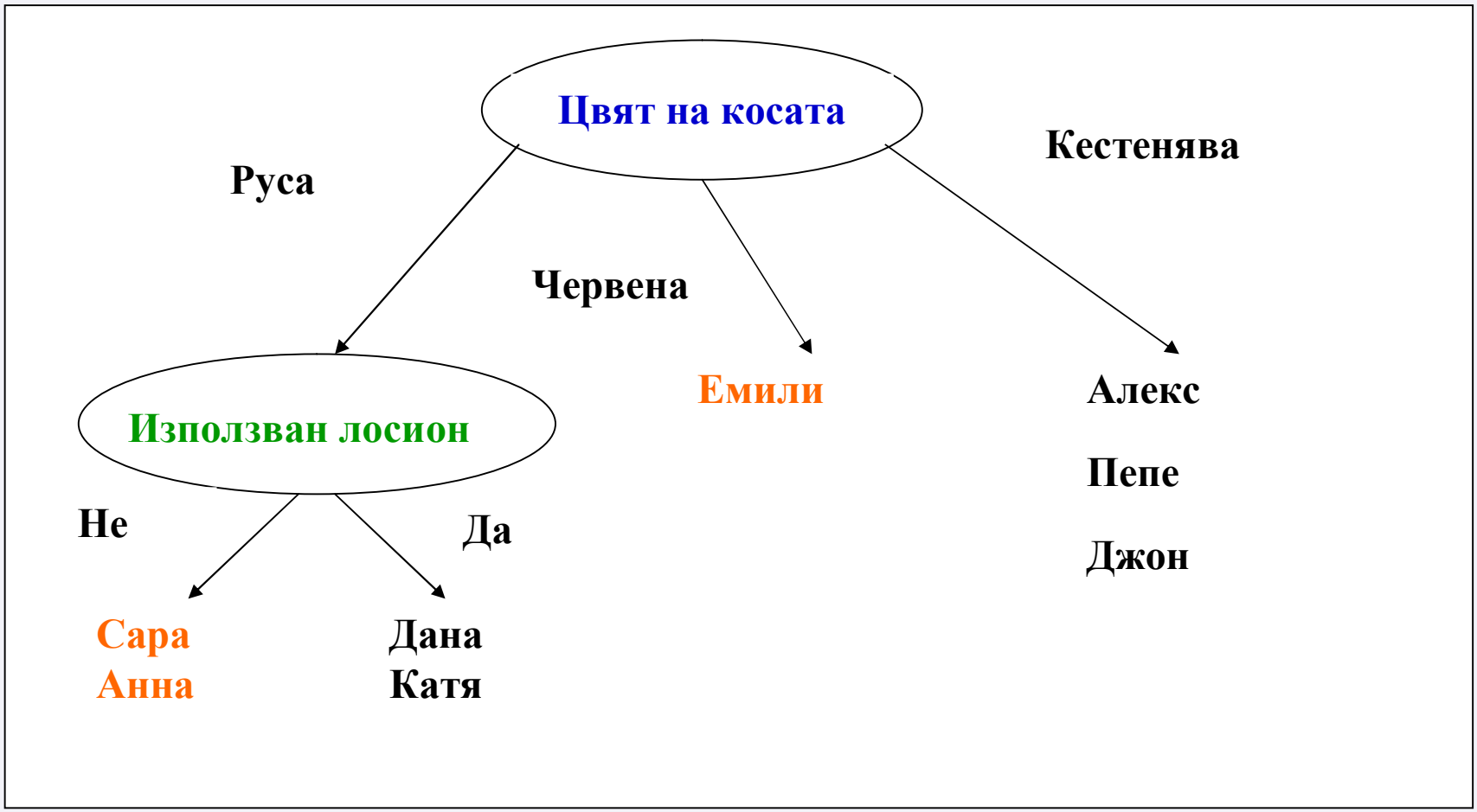
- Не може да се използва и метода пространство на версиите поради същата причина.

Построяване на идентификационно дърво.

Нека опитаме да представим нашата таблица във вид на граф, като във връх записваме свойство, а изходящите му дъги означим със стойности на това свойство.

Листата на дървото нека включват всички индивиди, които се характеризират със стойностите на свойствата, които са разположени по пътя от корена на дървото до конкретното листо.

Построяване на идентификационно дърво.



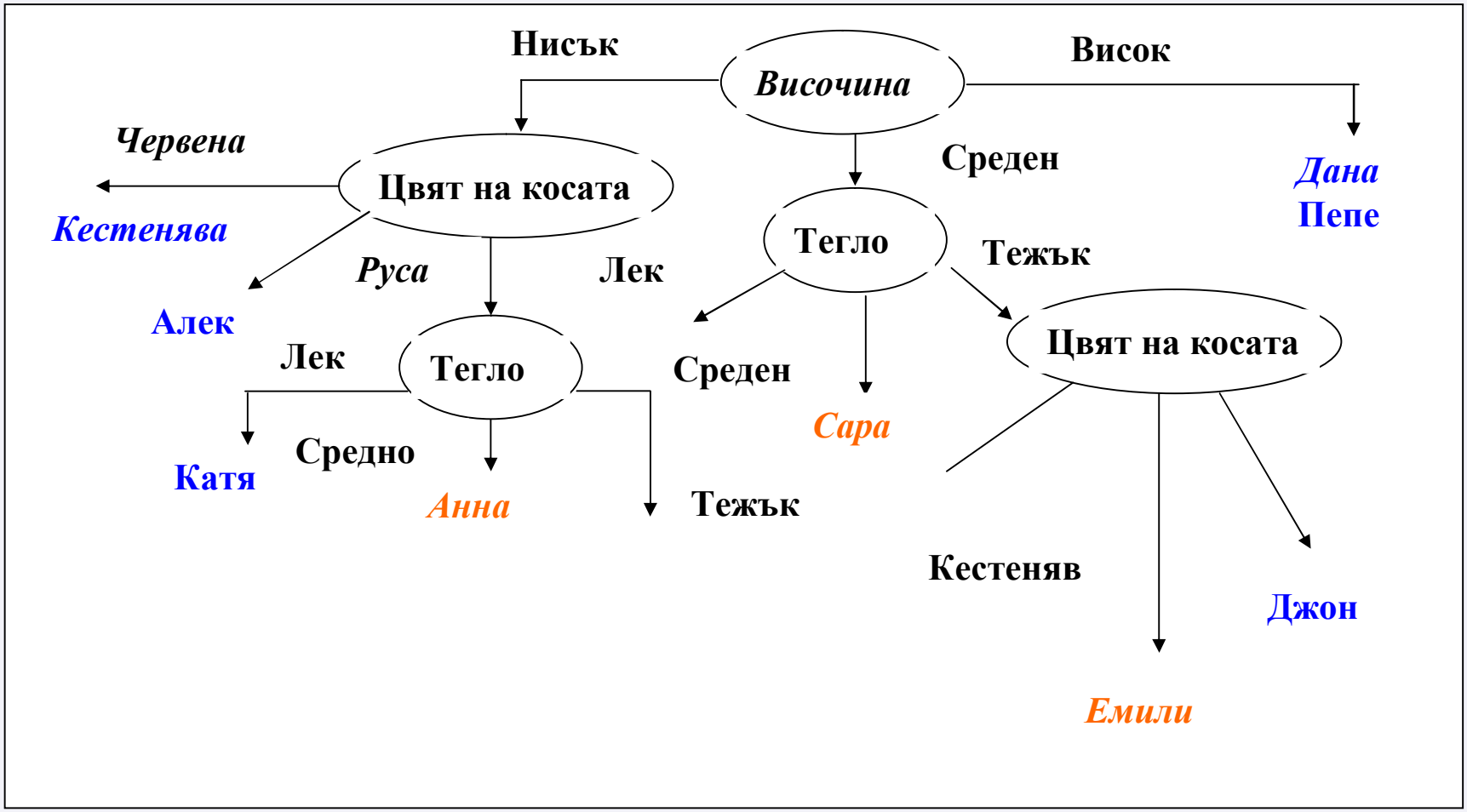
Построеното идентификационно дърво е добре подредено и изглежда завършено.

Идентификационното дърво е дърво на решенията, в което всяко подмножество от възможни заключения се основава върху списък от примери на познат клас.

Този начин за представяне на свойствата с техните стойности при подреждането на индивидите представлява подреждане в идентификационно дърво.

Съществено за него е, че за всеки индивид сме означили по подходящ начин интересуващото ни свойство – дали той ще изгори или не..

За построеното тук идентификационно дърво изходни са същите наблюдения, както при показаното преди, но то изглежда по съвсем различен начин и за него съвсем не може да се каже, че е подредено.



- Идентификационното дърво ще бъде завършено тогава, когато всяко листо съдържа само индивиди, които или изгарят на слънцето или не.
- Тогава идентификационното дърво е дърво на решенията или **решаващо дърво.**
- Попадането на нов индивид при спускане по дървото в конкретно листо еднозначно ще определи дали той ще изгори или не.

- Ние намерихме подход - процедура за тестване на свойствата, която да класифицира правилно всеки от примерите.

Когато такава процедура заработи правилно с достатъчно голям брой примери, може с увереност да се предположи,че тя ще може да работи и с обекти, чиято класификация е непозната.

- **Идентификационното дърво е решаващо дърво, в което всяко множество от възможни заключения е напълно установено чрез списък от примери на познати класове.**

Минимално идентификационно дърво.

- Но как може програмата да достигне до верни заключения без някакви предварителни знания за действията на лосиона или цвета на косата и как те се свързват със свойствата на кожата.

Отговор предлага следната евристика:

- Светът по принцип е прост. **Следователно най-малкото идентификационно дърво, което съответства на примерите, е това, което най-вероятно ще идентифицира непознатите обекти най-правилно.**

- Следователно въпросът се изменя

от кое е правилното идентификационно дърво

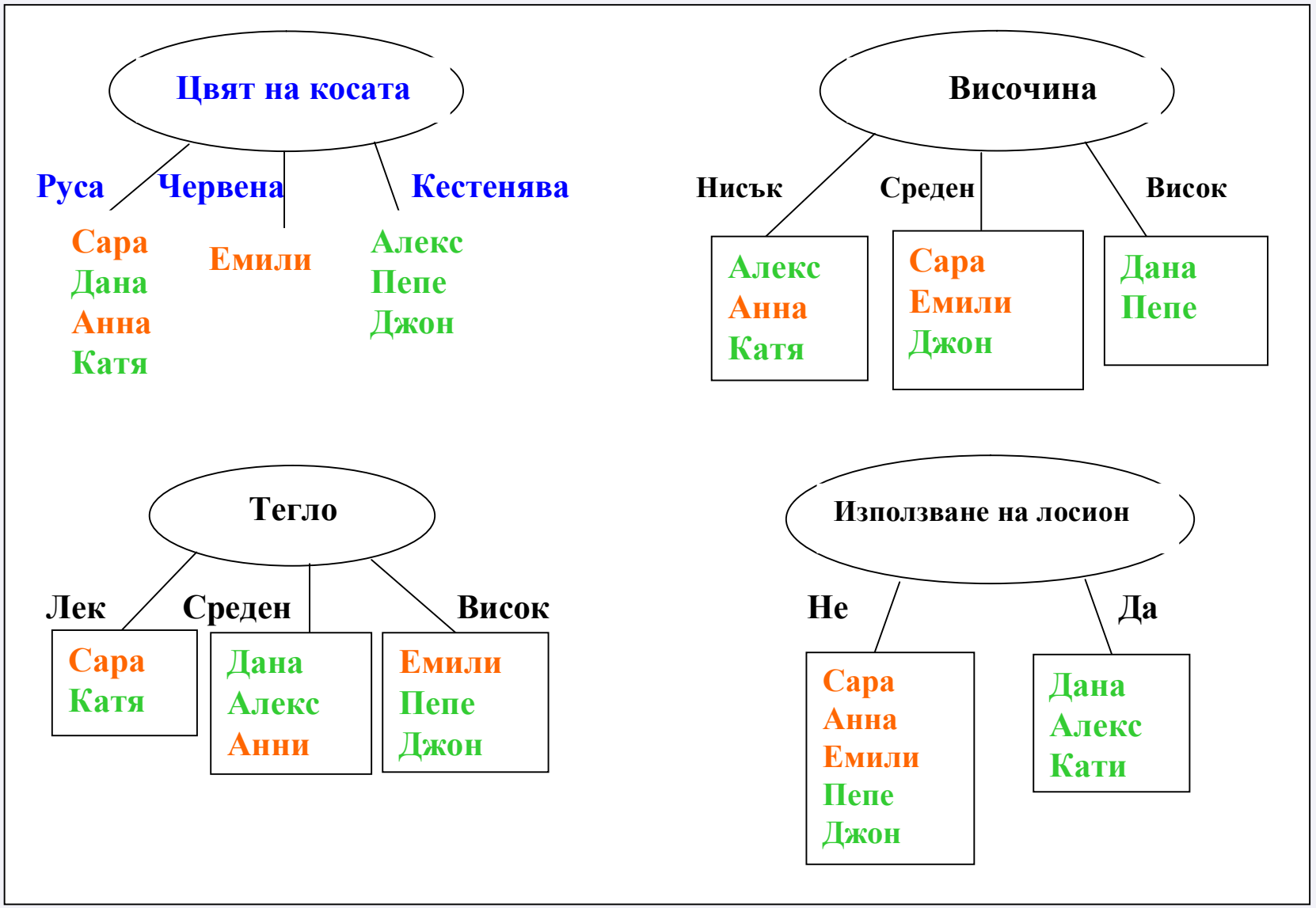
до въпроса как може да се конструира най-малкото идентификационно дърво.

- За съжаление тази задача е трудно решима при голям брой тестове, затова ще се задоволим с една процедура, която има тенденция да строи малки дервета без да ни гарантира, че построеното дърво е най-малкото възможно за случая.

- Водеща идея - **тестовете трябва да минимизират безредието.**

- Един начин за реализиране на минимално идентификационно дърво е да се избере за основа тест, който върши най-добра работа от гледна точка на разделянето на примерите в базата данни на подмножества, в които има най-голям брой примери от един единствен вид - в хомогенни подмножества.
- След това за всяко подмножество, което съдържа примери от различен вид, се избира друг тест, за да се раздели това нехомогенно подмножество на хомогенни подмножества

Разглеждаме отново примерната база данни за изгаряниятa от слънцето.



Както се вижда от фигурата

Тестът за теглото е най-неподходящия избор за основен тест, тъй като нито едно от подмножествата не е **хомогенно подмножество**. За да продължим са необходими три нови теста.

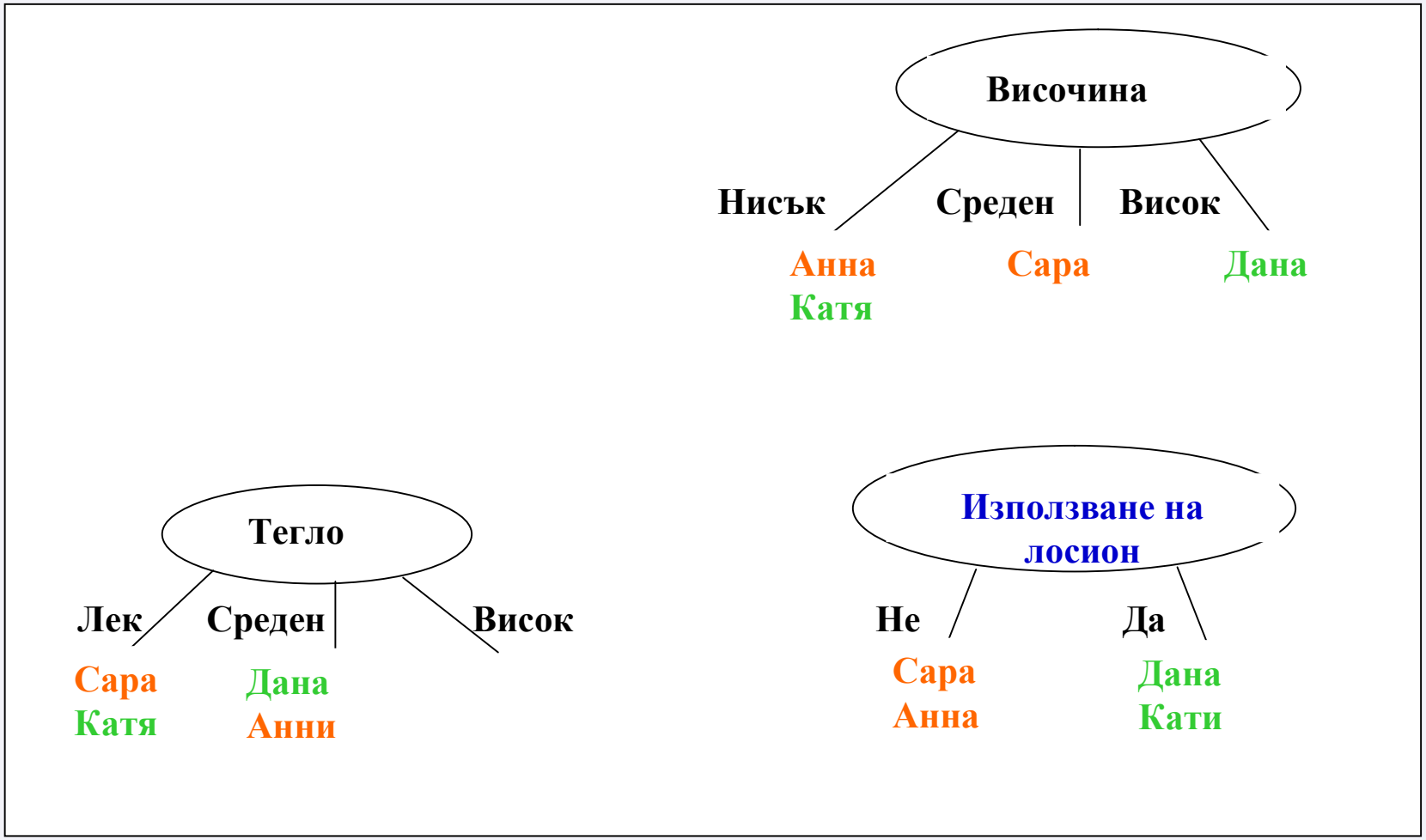
Тестът за височината е малко по-добър, защото за продължение са необходими два нови теста (**двама човека** са попаднали **в хомогенно подмножество -** *Дана и Пепе*).

Тестът за лосиона е още по-добър, защото продължението изисква един тест като **трима човека** са в **хомогенна подмрежа** /*Дана, Алекс и Катя*/.

Тестът за цвета на косата е най-подходящ да бъде избран за основен, защото продължението изисква един тест и **четири човека** попадат в **хомогенна подмрежа**: /*Емили, Алекс, Пепе и Джон*/.

Затова като първи се избира теста за цвят на косата. Този тест води само до една нехомогенна подмрежа състояща се от *Сара, Дана, Анна и Катя*.

За да разделим това подмножество по-нататък разглеждаме останалите три теста за тези четири човека от нехомогенната мрежа.



Измерване на неподредеността.

Теорията на информацията ни предлага формула за измерване на неподредеността.

За реална база данни с големи размери е твърде малко вероятно един тест да даде напълно хомогенни подмножества.

Следователно за реални бази данни е необходим полезен начин за измерване на цялата неподреденост или нехомогенност в подмножествата, получени от всеки тест.

Необходимата формула може да се вземе от теорията на информацията.

Формулата за неподреденост на едно множество

Формулата за неподреденост на едно множество **B,** състоящо се от $\mathbf{n_b}$ елемента, което съдържа **c** на брой хомогенни подмножества, всяко съответно с по $\mathbf{n_{bc}}$ елемента, е следната:

$$Неподреденост = \sum_c - \frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b}$$

За средната неподреденост на едно множество може да се използува следната формула:

$$Средна \quad неподр = \sum_b \frac{n_b}{n_t} * \left(\sum_c - \frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b} \right)$$

където: **b** е брой подмножества,

$\mathbf{n_b}$ - брой примери в подмножеството **b**;

$\mathbf{n_t}$ - общият брой от примери във всички подмножества;

$\mathbf{n_{bc}}$ -всички примери от подмножеството **b**, които са от клас **c**.

Тестът дава
най-голяма стойност на средната неподреденост, когато неподредеността е най-голяма, и
най-малка стойност, когато тестът води до напълно хомогенни подмножества.

Неподредеността варира от 0 до 1.

Неподредеността е 0, когато подмножествата са напълно хомогенни и неподредеността е 1, когато подмножествата са изцяло нехомогенни

След като разполагаме с начин за изчисляване на неподредеността в едно множество можем да измерваме средната неподреденост на множествата, получени като изход от всеки тест.
За да получим средната неподреденост на идентификационното дърво, необходимо е да се претегли неподредеността във всяко разклонение на множеството чрез размера на множеството, отнесено към пълния размер на всички разклонения в подмножествата.

За разглеждания пример се получават следните резултати:
Средната неподреденост(цвят на косата)=0,5;
Средната неподреденост(височина)=0,69;
Средната неподреденост(тегло)=0,94;
Средната неподреденост(лосион)=0,61.

Генериране на идентификационно дърво:

Докато всеки възел-листо не е съставен само от хомогенни подмножества:

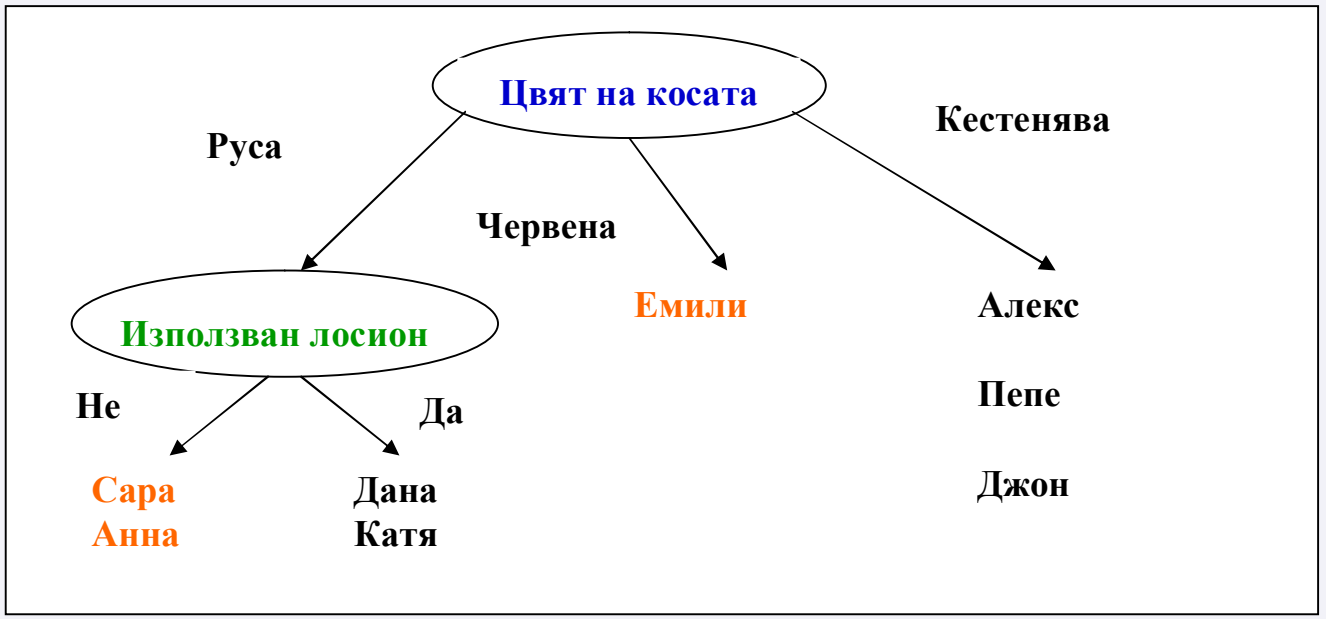
1)Избирай възел-листо с нехомогенно подмножество;

2)Замени това възел-листо с тестов възел, който разделя нехомогенните подмножества на минимални нехомогенни подмножества съгласно някакво измерване на неподредеността.

От дървета към правила.

След като идентификационното дърво е построено, то може сравнително просто да се преобразува в група от еквивалентни му правила.
***Проследяват се всички пътища в дървото от корен до листо.**
****Тестовете, през които се минава, се записват като предпоставки, а заключението е класът на елементите във върха-листо.**
За нашия пример:
Правило1.
1.Ако косата на индивида е руса.
2.Ако индивидът не използува лосион.
Извод2:Индивидът изгаря от слънцето.
Правило2.
1.Ако косата на индивида е руса.
2.Ако индивидът използува лосион.
Извод1:Нищо не се случва.
Правило3.
1.Ако косата на индивида е червена.
Извод2:Индивидът изгаря от слънцето.
Правило4.
1.Ако косата на индивида е кестенява.
Извод1:Нищо не се случва.

От дървета към правила.



Правило1.
1.Ако косата на индивида е руса.
2.Ако индивидът не използува лосион.
Извод2:Индивидът изгаря от слънцето.

Правило2.
1.Ако косата на индивида е руса.
2.Ако индивидът използува лосион.
Извод1:Нищо не се случва.

Правило3.
1.Ако косата на индивида е червена.
Извод2:Индивидът изгаря от слънцето.

Правило4.
1.Ако косата на индивида е кестенява
Извод1:Нищо не се случва.

Упростяване на множеството правила

След като е построено множеството от правила може да се опрости:

(1)Опитваме да опростим всяко правило като се елиминират ненужните предпоставки

(2)Опитваме се да отстраним ненужните правила.

Ненужните предпоставки(условия) трябва да се премахнат.

За да се опрости едно правило, се проверява дали някои от предпоставките му могат да се елиминират без това да промени разделянето на случаите.

- Например второто правило има две предпоставки.

Правило2.

1.Ако косата на индивида е руса.

2.Ако индивидът използува лосион.

Извод1:Нищо не се случва.

Ако елиминираме първата предпоставка, която проверява дали косата е руса, правилото ще се отнася до всеки човек, който използва лосион.

В дадените примери всички хора, които използват лосион, не изгарят от слънцето независимо от това какъв цвят на косата имат.

В случая отпадането на проверката за цвят на косата не води до промени в класификацията.

Това означава вероятно, че цветът на косата не е съществена предпоставка за това правило.

Пропускаме я и получаваме следното правило:

- **Правило2А.**
- **1. Ако индивидът използува лосион.**
- **Извод1:Нищо не се случва**

Ненужните предпоставки(условия) трябва да се премахнат.

- **Правило2А.**
- **1. Ако индивидът използува лосион.**
- **Извод1:Нищо не се случва**

За да станат тези разсъждения по-убедителни, полезно е да се построят таблици, които показват в каква степен резултатът зависи от свойството.

Да разгледаме следната таблица, в която е показано разделянето на използуващите лосион на русокоси и нерусокоси и съответно на изгоряли от слънцето и неигоряли (без промени).

Таблицата показва ясно, че в случай на използуване на лосион знанието дали индивидът е русокос или не е русокос, няма отношение към определянето дали ще изгори или не,.

Ползват лосион	Без промяна	Изгаря
Русокос	2	0
Не е русокос	1	0

Правило2.
1.Ако косата на индивида е руса.
2.Ако индивидът използува лосион.
Извод1:Нищо не се случва.

- **Правило2.**
- **1.Ако косата на индивида е руса.**
- **2.Ако индивидът използува лосион.**
- **Извод1:Нищо не се случва.**

Нека проверим дали може да се отстрани втората предпоставка в това правило.

Таблицата за русокосите ползващи и неползващи лосион показва, че <u>това условие определя еднозначно</u> дали русокосият човек ще изгори или не.

Втората предпоставка не може да се отстрани.

Русокос	Без промяна	Изгаря
Ползва лосион	2	0
Не ползва лосион	0	2

***Елиминиране на ненужни правила*.**

След като се опростят правилата като се елиминират някои от техните предпоставки, необходимо е да се опрости множеството от правила, като се отстранят излишните правила.

Това например са правилата, които съдържат в себе си изцяло текста на друго правило.

Елиминиране на ненужни правила.

Друг вариант за упростяване ще демонстрираме с разглеждания пример за слънчевото изгаряне, където получихме следните четири правила:

Правило1.

1.Ако косата на индивида е руса.

2.Ако индивидът не използува лосион.

Извод2:Индивидът изгаря от слънцето.

Правило2А.

1. Ако индивидът използува лосион.

Извод1:Нищо не се случва

Правило3.

1.Ако косата на индивида е червена.

Извод2:Индивидът изгаря от слънцето.

Правило4.

1.Ако косата на индивида е кестенява.

Извод1:Нищо не се случва.

- **Вижда се, че две правила (1 и 3) показват, че човек се връща изгорял от слънцето, и две правила (2А и 4), които показват, че човек няма изгаряния.**
- **Тогава може двете правила, които дават заключението, че човек се е върнал изгорял от слънцето да се заменят от подразбиращо се правило, което се използва, ако няма друго правило, което може да се използва.**

Елиминиране на ненужни правила.

В нашия пример може да се използуват следните правила

- Правило2А.

1. Ако индивидът използува лосион.

Извод1:Нищо не се случва

- Правило4.

1.Ако косата на индивида е кестенява.

Извод1:Нищо не се случва.

- Правило5.

1.Ако не може да се приложи друго правило.

Извод2:Индивидът изгаря от слънцето.

Процедура за генериране на правила от идентификационно дърво:

- 1)Създава се по едно правило за всеки път от корен до листо в идентификационното дърво.
- 2)Опростява се всяко правило чрез премахване на предпоставките, които не влияят върху изводите на правилото.
- 3)Правилата се групират по вида на изводите си и групата с най-голям брой включени правила се замества от подразбиращо се правило, което се използува, когато няма възможност да се използува друго правило

Прецизно тестване на правила

Нека спрем вниманието си на следната таблица, която ни показва
наличието (**R**) или отсъствието(¬**R**) на резултата **R** в зависимост
от наличието(**P**) или отсъствието(¬**P**) на свойството **P**.

	R	¬**R**
P	k	m
¬**P**	n	r

- **В каква степен можем от знание за наличие или отсъствие на причината да правим извод за наличие или отсъствие на резултата и зависи ли тази степен от стойностите в тази таблица (таблица на случайната извадка)?**

Прецизно тестване на правила

Нека тестваме дали да отстраним предпоставката **P** в правило с резултат **R** и получим за четворката (**k, m, n, r**) стоиности 1,0,01.

Очевидно предпоставката е важна, тъй като отстраняването ѝ води до игнориране на половината от случаите.

	R	**¬R**
P	k	m
¬P	n	r

Ако тестовата извадка се състои от 1000 случая и съдържанието на таблицата е 999,0,0,1, отстраняването на предпоставката ще пренебрегне само един от случаите, или с опасност от допускане на грешка 1 на хиляда бихме могли да си позволим да отстраним предпоставката от правилото.

С колкото повече данни разполагаме, толкова разсъжденията ни ще са по-близо до истината.

При повече данни появата на случайни грешки в наблюденията би повлияло по-незначително върху резултатите.

Ако числата са малки, за препоръчване е да се откажем от предпоставката, тъй като за нейното използуване нямаме достатъчно потвърждения.

Статистическа зависимост

- **А как стои въпросът с относителните величини k/m и n/r ?**

- **Ако отношението k/m е същото или почти същото като n/r знанията относно P едва ли са полезни и е по-добре да се откажем от предпоставката.**

- **Същевременно, ако броят на наблюденията е голям и отношението k/m силно се различава от n/r, знанията относно P <u>са много показателни и предпоставката обезателно трябва да остане.</u>**

	R	¬R
P	k	m
¬P	n	r

Статистическа зависимост

- **Да скицираме няколко стъпки, които статистическата теория ни препоръчва, за да достигнем до прецизни изводи относно последното твърдение.**

- Нашата цел е да установим дали съществува статистическа зависимост между предпоставката и извода.

- Тъй като статистическата зависимост може да се проявявя в много форми и ние не знаем предварително за коя да проверяваме, а **статистическата независимост има само една форма**, по-лесно е да се установи обратното, т.е. липсата на статистическа зависимост между предпоставката и извода.

- *Вместо да доказваме, че резултатът **R** зависи от предпоставката **P**, ще се опитваме да покажем, че не е правдоподобно резултатът **R** да не зависи от предпоставката **P**.*

Маргинални суми

Нека допълним таблицата с т.н. маргинални суми, т.е. суми по редове и суми по колони.

	R	**¬R**	**Суми**
P	k	m	**SP =k+m**
¬P	n	r	**S¬P =n+r**
Суми	**SR=k+n**	**S¬ R =k+n**	**SP+S¬P = SR+S¬ R**

Забележете, че ако са ви известни тези суми, то само по едно от числата в таблицата бихте могли да възстановите останалите.

Нека ни е известна стойността на величината **k.** Време е да приведем една грандиозна формула, която ни показва каква е вероятността величината **k** да приеме конкретна стойност в зададен интервал от стойности.

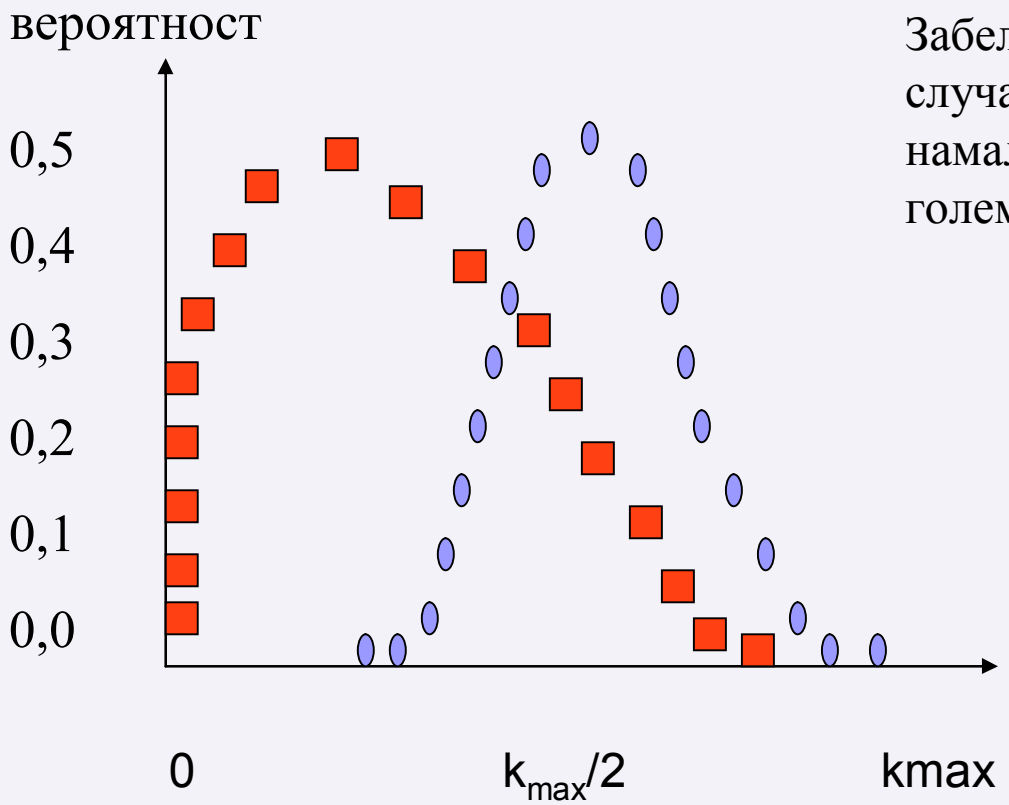
Формулата, която ни показва каква е вероятността величината k да приеме конкретна стойност в зададен интервал от стойности.

$$P(k/SP, S\neg P, SR, S\neg R) = \frac{\frac{SP!}{k!(SP-k)!} \times \frac{S\neg P!}{(SR-k)!(S\neg P-(SR-k))!}}{\frac{(SP+S\neg P)!}{SR!(SP+S\neg P-SR)!}}$$

Във формулата $S\neg R$ не се появява, защото е представена чрез останалите величини.

Ако построим зависимостта между изчислената вероятност и стойностите за **k** при **SP = S¬P =SR= S¬ R,** ще получим синята крива от фигурата.

В случай на неравество между **SP** и **S¬P** кривата ще се отмести, както е показано на същата фигура с червени квадратчета, като симетрията ще се наруши.



Забележете, че при симетричния случай вероятностите бързо намаляват към малките и към големите стойности за **к**.

Следователно вероятността стойността на **к** да е извън централната област на кривата е много малка.

Това означава, че предпоставката и резултата са независими

- При симетричния случай вероятностите бързо намаляват и към малките, и към големите стойности за **к.**
- Следователно вероятността стойността на **к** да е извън централната област на кривата е много малка.
- Това означава, че предпоставката и резултата са независими.
- *С други думи вероятостта резултатът да е зависим от предпоставката е нищожно малка и попада в статистическата грешка*.

- **Следователно статистическа зависимост е налице тогава, когато числата SP и S¬P се различават значително.**

- **В случаите, когато тези две числа са близки по стойност най вероятно резултатът и предпоставката са статистически независими и в правилото предпоставката трябва да се изостави**

Заключение:

- Компютрите могат да се обучават чрез формиране на предварително събрана информация във вид на идентификационно дърво.
- Светът по принцип е прост.
- Следователно най-малкото идентификационно дърво,

което съответства на примерите, най-добре ще идентифицира непознатите обекти.

- Възможен начин да се построи минимално идентификационно дърво е да се използува формулата за неподредеността, заимствана от теория на информацията.
- Идентификационното дърво може да се превърне в система от правила, което прави придобитото знание по-удобно за използуване. Преобразуването се осъществява като всеки път в дървото се превръща в правило, тестовете по този път – в условия на правилото, а същността на листото – в следствие на правилото.
- Правилата могат да се упростят като най-напред се елиминират излишните условия, а след това се елиминират излишните правила.
- Важно изискване при използуването на този подход е да се убедим, че между условие и следствие съществува статистическа зависимост.

Литература.

[Quinlan, R., Discovering Rules by Induction from Large Collections of Examples, in Exper Systems in the Microelectronic Age, Edinburgh University Press, Edinburgh, Scotland, 1979]

– За първи път се описват идентификационните дървета.

[Quinlan, R.,Simlifying Decision Trees, Report HM-930, *Artificial Intelligence Laboratory*, MIT, Cambridge, MA, 1986]

– Описват се възможностите от идентификационните дървета да се извличат правила.

[Quinlan, R., R. Divest, Inferring Decision Trees Using the Minimum Description Length Principle, Report TM-339**,** Laboratory of Computer Science, MIT, Cambridge, MA, 1987]

– Описва се процедура за идентификация, която използува минимум памет.

[Dakovski, L., Z. Shevked, A novel approach for building identification tree from examples, Proc. of the Second International Scientific Conf. on Computer Science'2005, Proc. Part I, Greece, pp 214-219.]

- Описва се процедура за построяване на минимално идентификационно дърво