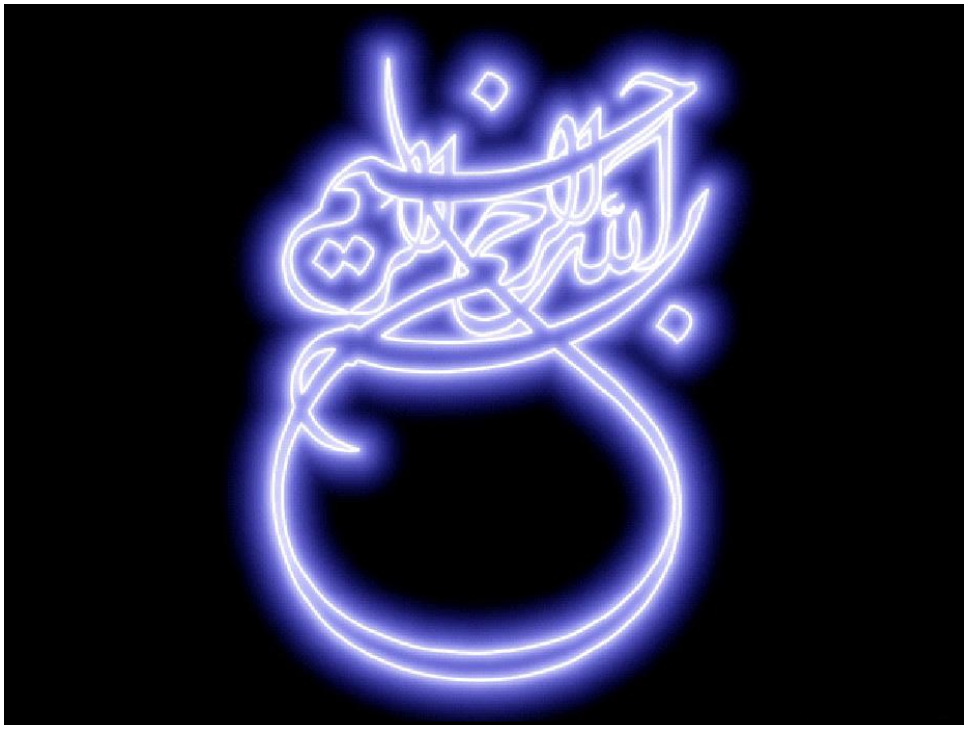
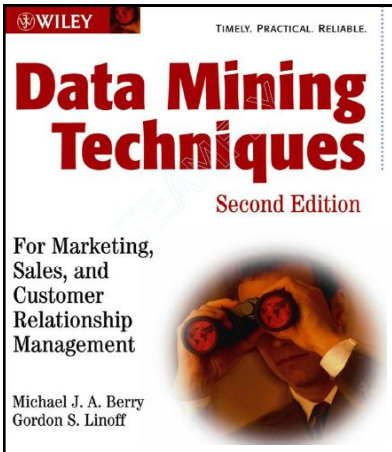
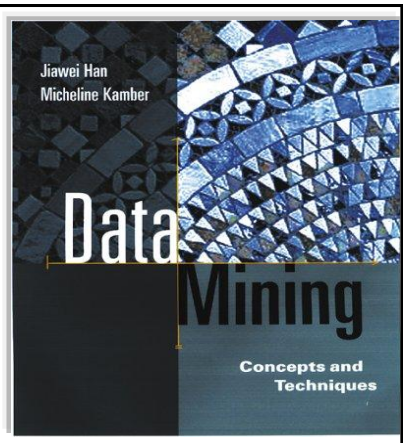






داده کاوی
دسته بندی
درخت تصمیم

سمیه علیزاده

درخت تصمیم گیری

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

تعریف

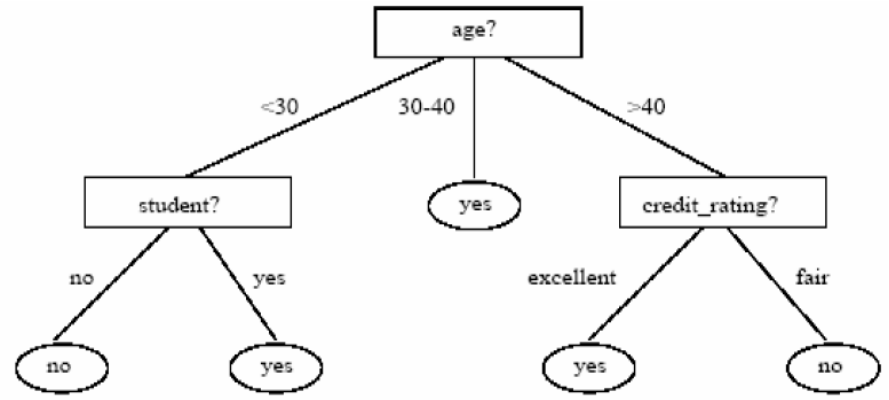
- درخت تصمیم یکی از ابزارهای قوی و متداول برای دسته بندی و پیش بینی می باشد.
- درخت تصمیم گیری یک ساختار درختی شبیه فلوچارت است.
- در این ساختار هر گره داخلی آزمونی را بر روی یک ویژگی مشخص می کند.

گره های برگ، کلاسها یا توزیع کلاسها را ارائه می نمایند.

- بالاترین گره در درخت گره ریشه است.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

ساختار درخت تصمیم گیری



شکل 1 . نمونه ای از یک درخت تصمیم گیری خرید کامپیوتر در AllElectronics

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

خصوصیات درخت تصمیم گیری

- درخت تصمیم پیش بینی خود را در قالب یکسری قوانین توضیح می دهد در حالیکه در شبکه های عصبی تنها پیش بینی بیان می شود و چگونگی آن در خود شبکه پنهان باقی می ماند.
- همچنین در درخت تصمیم گیری بر خلاف شبکه های عصبی لزومی ندارد که داده ها لزوما بصورت عددی باشند.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

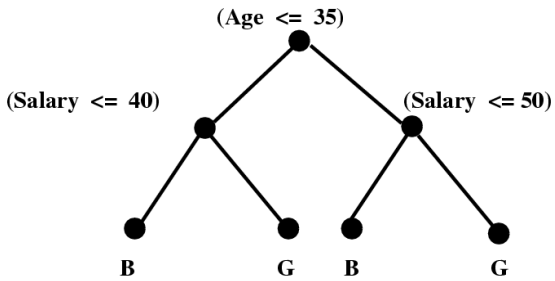
درخت تصمیم گیری چگونه کار می کند؟

- در درخت تصمیم گیری نیز یکسری سوال وجود دارد و با مشخص شدن پاسخ هر سوال یک سوال دیگر پرسیده می شود. اگر سوالها درست وخوب پرسیده شوند یکسری کوتاه از سوالات برای پیش بینی دسته رکورد جدید کافی می باشد.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مثال

Age	Salary	Class
30	65	G
23	15	B
40	75	G
55	40	B
55	100	G
45	60	G



Classification rules:
Class B: (Age <= 35 AND Salary <= 40) OR (Age > 35 AND Salary <= 50)
Class G: (Age <= 35 AND Salary > 40) OR (Age > 35 AND Salary > 50)

Test data: Age = 25 AND Salary = 50
Class = G

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

اثر بخشی یک درخت تصمیم گیری

- درصد داده هایی که درست دسته بندی می شوند و دسته پیش بینی شده با دسته واقعی آنها یکسان است.
- همچنین کیفیت شاخه های ایجادشده نیز مهم است. هرراه ایجادشده از ریشه به یک برگ معادل یک قانون است که بعضی قانونها بهتر از سایر قانونها می باشند. در بعضی اوقات بریدن برخی شاخه های ضعیف تر درخت باعث بهبود قدرت پیش بینی درخت می شود.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Training Database

- مهمترین هدف از دسته بندی (Classification & Regression) بدست آوردن مدلی برای پیش بینی می باشد. بدین منظور از مجموعه ای به نام داده های آموزشی (Training Database) که مجموعه ای از متغیرها و رکوردها است استفاده می کنیم.

مثال:

Age	Car Type	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

انواع متغیر های درخت تصمیم گیری

- متغیر های عددی Numerical مانند سن، قد
- متغیر های رده ای Categorical مانند نوع، جنس
- از این متغیر ها برای پیش بینی متغیر هدف یا متغیر وابسته استفاده می کنیم.
- Predictor Attributes: به متغیر های گفته شده در مثال قبل (Age and Car type) متغیر های مستقل می گویند. و با گره ها نشان می دهند.
- Class Label: به متغیر های وابسته می گویند و با برگ نشان می دهند. در مثال قبل Risk.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Classification & Regression

- اگر متغیر وابسته از نوع عددی باشد مسأله به یک مسأله Regression تبدیل خواهد شد.
- و اگر رده ای باشد مسأله Classification است.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مراحل ایجاد درخت تصمیم گیری

- مرحله رشد و ایجاد درخت
- مرحله هرس درخت (که هدف این مرحله کاهش خطا ها می باشد.

الگوریتمهای متفاوتی برای ایجاد درخت وجود دارد

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

الگوریتم کلی ایجاد درخت

```
1-apply ss to D to find the splitting criterion
2-if n split
3-use best split to partition D to D1 and D2
4-Build tree(n1,D1,ss)
5-Build tree(n2,D2,ss)
6-end if
```

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

متدهای انتخاب نقطه شکست شاخه
Split Selection

- (gini(T) = 1 - Σpj2): Gini Index
- (entropy(T) = - Σpj x log2(pj)): Entropy
- Cart
- 2pj
- Min(pj)
- C4.5

که pj فراوانی نسبی از کلاس j در درخت T می باشد.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Gini Index

- در روش Gini Index همه متغیر ها را در گره امتحان کرده و آن متغیری که از همه کوچکتر باشد را انتخاب می کنیم.
- بهترین انتخاب برای تقسیم مجموعه S به دو مجموعه S1 و S2 از معیار زیر تبعیت می کند، یعنی ماکزیمم کردن تابع زیر:

$$I(S)-|S1|/|S|*I(S1)+|S2|/|S|*I(S2)$$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Gini Index مثال

Age	Car Type	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Gini Index مثال

ابتدا جدول را بر اساس متغیر age به صورت صعودی مرتب می کنیم.

Age	Car Type	Risk
17	Sports	High
20	Family	High
23	Family	High
32	Truck	Low
43	Sports	High
68	Family	Low

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

- حال از متد Gini Index برای انتخاب انشعاب استفاده می کنیم. هر دو متغیر Age و Car Type را بررسی می کنیم. اما قبل از اختصارات را معرفی می نماییم:

H: High

L: Low

R: Right Child

L: Left Child

$Gini(T)=1-\sum pj^2$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

- Age<=17

	H	L
L	1	0
R	3	2

$I(S1): 1-(1/1)^2-(0/1)^2 = 1\text{-}1\text{-}0=0$

$I(S2): 1-(3/5)^2-(2/5)^2 = 1\text{-}9/25\text{-}4/25=0.48$

$|s1|=1, |s2|=5, |s|=6 \Rightarrow I(S): |1|/|6|*0+ |5|/|6|*0.48 = 0+ 5/6*0.48=0.4$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Age<=20 •

	H	L
L	2	0
R	2	2

$I(S1): 1-(2/2)^2-(0/2)^2 = 1\text{-}1\text{-}0=0$

$I(S2): 1-(2/4)^2-(2/4)^2 = 1\text{-}4/16\ \text{-}4/16=0.5$

$|s1|=2,\ |s2|=4, |s|=6 \Rightarrow I(S): |2|/|6|*0+\ |4|/|6|*0.5= 4/6*0.5=0.33$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Age<=23 •

	H	L
L	3	0
R	1	2

$I(S1): 1-(3/3)^2-(0/3)^2 = 1\text{-}1\text{-}0=0$

$I(S2): 1-(1/3)^2-(2/3)^2 = 1\text{-}1/9\text{-}4/9=0.444$

$|s1|=3,\ |s2|=3, |s|=6 \Rightarrow I(S): |3|/|6|*0+\ |3|/|6|*0.444= 3/6*0.444=0.222$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Age<=32 •

	H	L
L	3	1
R	1	1

$I(S1): 1-(3/4)^2-(1/4)^2 = 1-9/16-1/16=0.375$

$I(S2): 1-(1/2)^2-(1/2)^2 = 1-1/4-1/4=0.5$

$|s1|=4, \quad |s2|=2, |s|=6 \Rightarrow \quad I(S): \quad |4|/|6|*0.375 + |2|/|6|*0.5 =$

$4/6*0.375+2/6*0.5=0.4166$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Age<=43 •

	H	L
L	4	1
R	0	1

$I(S1): 1-(4/5)^2-(1/5)^2 = 1-16/25-1/25=0.32$

$I(S2): 1-(0/1)^2-(1/1)^2 = 1-0-1=0$

$|s1|=5, |s2|=1, |s|=6 \Rightarrow I(S): |5|/|6|*0.32 + |1|/|6|*0 = 5/6*0.32 +0=0.266$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

- Age<=68

	H	L
L	4	2
R	0	0

$I(S1): 1-(4/6)^2-(2/6)^2 = 1-16/36-4/36=0.444$

$I(S2): 1-(0/0)^2-(0/0)^2 = 1-0-0=1$

$|s1|=6, |s2|=0, |s|=6 \Rightarrow I(S): |6|/|6|*0.32 + |0|/|6|*1 = 6/6*0.444 +0=0.444$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

- برای بررسی متغیر های رده ای categorical و به منظور سهولت در انجام کار، جدول فراوانی های هر کلاس را از روی همان جدول اولیه برای متغیر های رده ای تشکیل داده و سپس محاسبات را مشابه همان روش قبل انجام می دهیم.

Car Type	High	Low
sports	2	0
Family	2	1
Truck	0	1

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Car type={sports}

	H	L
Car type = sports	2	0
Car type # sports	2	2

$I(S1): 1-(2/2)^2-(0/2)^2 = 1-1-0=0$

$I(S2): 1-(2/4)^2-(2/4)^2 = 1-4/16-4/16=0.5$

$|s1|=2, |s2|=4, |s|=6 \Rightarrow I(S): |2|/|6|*0 + |4|/|6|*0.5 = 4/6*0.5 +0=0.333$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Car type={family}

	H	L
Car type = family	2	1
Car type # family	2	1

$I(S1): 1-(2/3)^2-(1/3)^2 = 1-4/9-1/9=0.444$

$I(S2): 1-(2/3)^2-(1/3)^2 = 1-4/9-1/9=0.444$

$|s1|=3, |s2|=3, |s|=6 \Rightarrow I(S): |3|/|6|*0.444 + |3|/|6|*0.444 = 0.444$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Car type={truck}

	H	L
Car type = truck	0	1
Car type # truck	4	1

$I(S1): 1-(0/1)^2-(1/1)^2 =$ 1-0-1=0

$I(S2): 1-(4/5)^2-(1/5)^2 =$ 1-16/25-1/25=0.32

$|s1|=1, |s2|=5, |s|=6 \Rightarrow$ I(S): |1|/|6|*0 + |5|/|6|*0.32 = 5/6*0.32 +0=0.266

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

- پس از بررسی کلیه حالتها، مینیمم I(S) را بدست می آوریم:

Min{0.4,0.33,0.222,0.4166,0.266,0.444,0.303, 0.266,0.444}=0.222

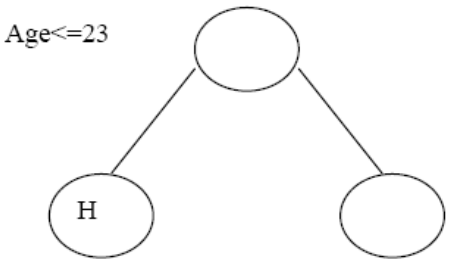
پس معیار Age<=23 را به عنوان نقطه انشعاب انتخاب می کنیم:

Age<=23 = {17,20,23}

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Age<=23={17,20,23}

Age	Car Type	Risk
17	Sports	High
20	Family	High
23	Family	High



چون دسته Class این مجموعه Label همه High می باشد:

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

- حال جدول Age>23 را تشکیل می دهیم تا معیار انشعاب بعدی با استفاده از همان روش فوق مجدداً برای این بخش از داده ها انتخاب شود:

Age	Car Type	Risk
32	Truck	Low
43	Sports	High
68	Family	Low

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Car type ={sports}

	H	L
Car type = sports	1	0
Car type # sports	0	2

$I(S1): 1-(1/1)^2-(0/1)^2 = 1\text{-}0\text{-}1=0$

$I(S2): 1-(0/2)^2-(2/2)^2 = 1\text{-}0\text{-}1=0$

$|s1|=1, |s2|=2, |s|=3 \Rightarrow I(S): |1|/|3|*0 + |2|/|3|*0 = 0$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Car type ={truck}

	H	L
Car type = truck	0	1
Car type # truck	1	1

$I(S1): 1-(0/1)^2-(1/1)^2 = 1\text{-}0\text{-}1=0$

$I(S2): 1-(1/2)^2-(1/2)^2 = 1\text{-}1/4\text{-}1/4=0.5$

$|s1|=1, |s2|=2, |s|=3 \Rightarrow I(S): |1|/|3|*0 + |2|/|3|*0.5 = 0.333$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Car type={family}

	H	L
Car type = family	0	1
Car type # family	1	1

$I(S1): 1-(0/1)^2-(1/1)^2 =$ 1-0-1=0

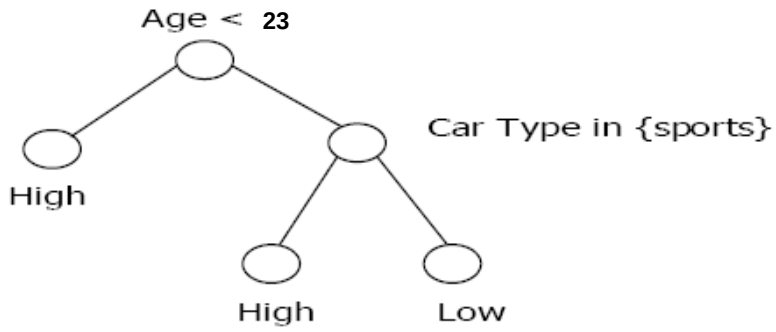
$I(S2): 1-(1/2)^2-(1/2)^2 =$ 1-1/4-1/4=0.5

$|s1|=1, |s2|=2, |s|=3 \Rightarrow I(S): |1|/|3|*0 + |2|/|3|*0.5 = 0.333$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

تکمیل درخت

- پس از بررسی تمام حالات مینیمم I(S) بین آنها صفر می باشد. پس درخت به شکل زیر تکمیل می گردد. این درخت نهایی است زیرا تمام برگهای آن به Class Label ختم شده اند.



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

ارزیابی درخت ایجاد شده

- برای محاسبه نرخ خطا در درخت ابتدا باید نرخ خطا در هر شاخه را بدست آوریم. نرخ خطا در هر برگ عبارتست از نسبت تعداد رکورد هایی که کلاس یا دسته آنها درست انتخاب یا پیش بینی نشده است.
- برای محاسبه خطای کل درخت، مجموع وزنی نرخ خطاهای برگ ها را بدست می آوریم.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

کیفیت درخت

- اگر مثلاً هدف دسته بندی بر اساس قد افراد باشد، و مجموعه ای ۱۱ نفری داشته باشیم که همه بجز یک محمد دارای قد کوتاه هستند، اگر این گره را به دو شاخه تقسیم کنیم، ممکن است قانونی به شکل زیر حاصل شود:
- "افراد کمتر از ۲۸ سال که نام آنها محمد است، بلند قد هستند.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

هرس کردن Pruning

- برای جلوگیری از چنین قانون هایی در بعضی از شاخه ها که شرایط خاصی در آنها وجود دارد، عملیات هرس با برش (Pruning) صورت می گیرد.
- این کار با آنکه نرخ خطا را افزایش می دهد ولی از ایجاد بعضی قانون های ناکارآمد جلوگیری می کند.
- البته نرخ خطای بدست آمده در درخت جدید نباید تفاوت چندانی با قبلی داشته باشد.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

نقاط قوت درخت تصمیم گیری

1- درخت تصمیم گیری به ما این توانایی را می دهد که پیش بینی های خود را در قالب یکسری قوانین ارائه دهیم .

2- درخت تصمیم گیری نیاز به محاسبات خیلی پیچیده ای برای دسته بندی داده ها ندارد.

3- درخت تصمیم گیری برای انواع مختلف داده ها از قبیل پیوسته و رده ای قابل استفاده می باشد.

4- درخت تصمیم گیری به ما نشان می دهد که کدام فیلد یا متغیرها تاثیرات مهمی در پیش بینی و دسته بندی ما دارند.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

نقاط ضعف درخت تصمیم گیری

1- بعضی از روشهای درخت تصمیم گیری تنها می توانند در مورد متغیرهای هدف دوتایی (بله یا خیر- پذیرش یا عدم پذیرش) دسته بندی و پیش بینی انجام دهند و در بعضی از آنها هنگامی که تعداد مثالهای هر کلاس کم باشد نرخ خطا بالا می رود.

2- این الگوریتم به حافظه زیادی نیاز دارد. در هر گره برای مقایسه فیلدها و محاسبه بهترین فیلد نیاز به بخاطر سپردن وضعیت هر فیلد می باشد که این حافظه زیادی نیاز دارد. همچنین در قسمت برش شاخه ها نیز برای انتخاب بهترین زیر درختی که می توان برش داد وضعیت هر زیر شاخه را بایستی بخاطر سپرد.

3- اکثر الگوریتمهای درخت تصمیم گیری در هر گره تنها یک فیلد را برای شاخه زدن در نظر می گیرند، در حالیکه ممکن است متغیرهای ترکیبی بهتر باشند