دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر

گزارش تکلیف دوم درس یادگیری ماشین

آشنائی با درخت تصمیم و ابزار weka

دانشجو:
سید احمد نقوی نوزاد
ش-د: ۹۴۱۳۱۰۶۰

استاد:
دکتر ناظرفرد

بهار ۹۵

سوال اول

الف) آنتروپی مجموعه بر اساس متغیر هدف (متغیر آخر هفته)

نکته‌ی قابل توجه در این سوال آن است که «هفته» در این‌جا در واقع یک ویژگی به حساب نیامده و حکم یک killer feature را دارد؛ اما از آن‌جائی که حکم زمان ثبت داده را دارد می‌تواند حائز اهمیت باشد، چون با توجه به این‌که تعداد ۳ داده‌ی آموزشی تکراری داریم (داده‌های ۶، ۹ و ۱۰ تکرار شده‌اند)، می‌توان با توجه به تفاوت زمانی ثبت آن‌ها، از حذفشان صرف نظر کرده و ثبت مجدد آن‌ها را به عنوان یک وزن برای داده‌های مربوطه به حساب آورد. لذا در این‌صورت وضعیت متغیر هدف در مجموعه‌ی داده‌های آموزشی مسئله به قرار زیر خواهد بود:

[6 Cinema, 2 Tennis, 1 Lesson, 1 Shopping]

حال برای محاسبه‌ی آنتروپی این مجموعه بر اساس متغیر هدف (متغیر آخر هفته) داریم:

$$Entropy\,(S) = -\frac{6}{10}\log^{\frac{6}{10}} - \frac{2}{10}\log^{\frac{2}{10}} - \frac{1}{10}\log^{\frac{1}{10}} - \frac{1}{10}\log^{\frac{1}{10}} = 1.5710 \leftarrow \text{Entropy before split}$$

ب) بهره‌ی اطلاعات ویژگی‌ها

Feature	Feature values distribution		Entropy	Information Gain	Split Entropy	GainRatio
Quiz	No	[5C, 0T, 0L, 0S]	0	InfoGain(S, quiz)= 1.5710-(.5*0+.5*1.9219)= **.6100**	**1**	$\frac{InfoGain(S,quiz)}{SplitEntropy(S,quiz)} = \frac{.6100}{1} = .6100$
	Yes	[1C, 2T, 1L, 1S]	1.9219			
Financial State	Bad	[3C, 0T, 0L, 0S]	0	InfoGain(S, financial state)= 1.5710-(.3*0+.7*1.8424)= **.2813**	**.8813**	$\frac{InfoGain(S,finState)}{SplitEntropy(S,finState)} = \frac{.2813}{.8813} = .3192$
	Good	[3C, 2T, 1L, 1S]	1.8424			
Climate	Sunny	[1C, 2T, 0L, 0S]	.9183	InfoGain(S, quiz)= 1.5710- (.3*.9183+.4*.8113+.3*.9183) = **.6955**	**1.5710**	$\frac{InfoGain(S,climate)}{SplitEntropy(S,climate)} = \frac{.6955}{1.5710} = .4427$
	Stormy	[3C, 0T, 0L, 1S]	.8113			
	Rainy	[2C, 0T, 1L, 0S]	.9183			

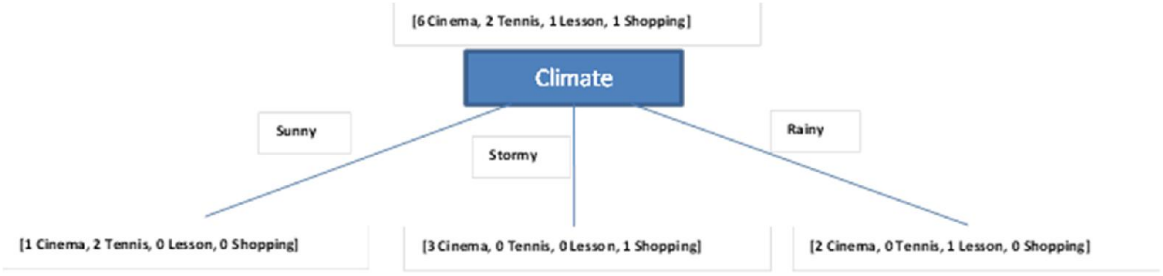
ج) انتخاب ریشه‌ی درخت تصمیم

در اینجا باید آن ویژگی را به عنوان ریشه‌ی درخت تصمیم انتخاب نمائیم که اندازه‌ی معیار مربوطه برای آن ویژگی نسبت به سایر ویژگی‌ها بیشینه باشد.

Feature selection criterion	Selected feature as the Root of DT	Misclassified items	Conclusion
InfoGain	Climate	0	Overfitting
GainRatio	Quiz	0	Overfitting

همانطور که پیداست هیچ داده‌ای به اشتباه دسته‌بندی نشده و این نشان از overfitting الگوریتم درخت تصمیم بر روی داده‌های آموزشی ما دارد.

درخت حاصل تا عمق یکم با توجه به معیار InfoGain به صورت زیر خواهد بود:



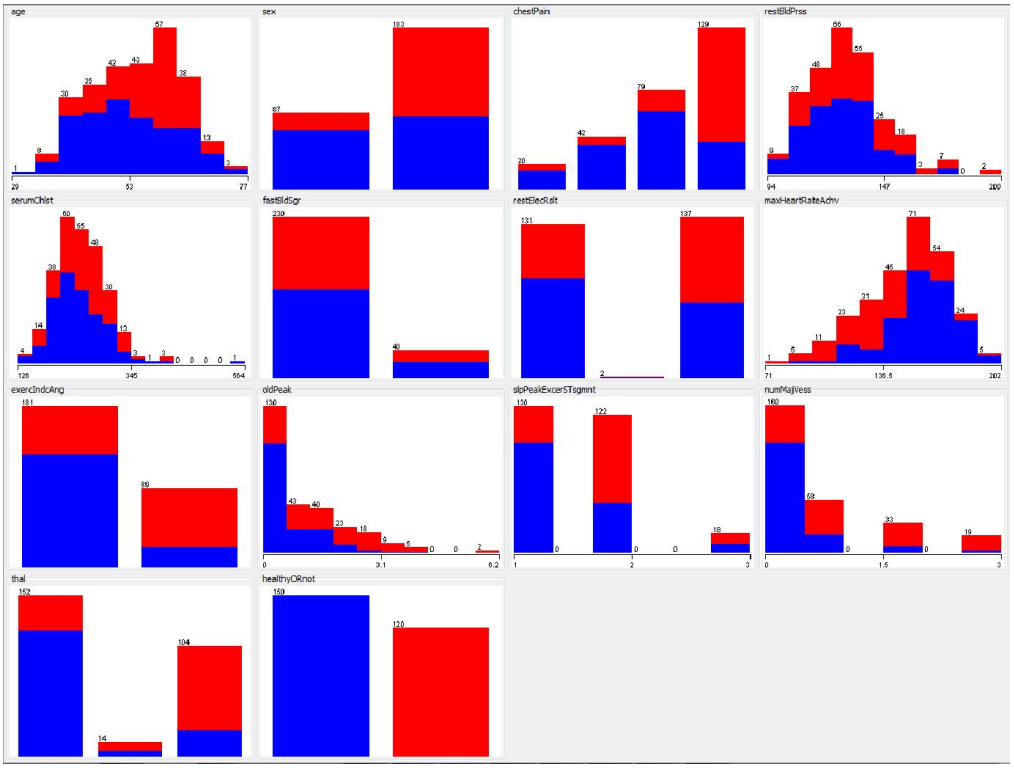
سوال دوم

بخش اول: آماده‌سازی مجموعه داده

در اینجا باید با توجه به مقادیر و عنوان هر ویژگی، عددی (numerical) یا اسمی (nominal) بودن آن را تعیین نمود و با توجه به اینکه پیش‌فرض weka برای داده‌های ورودی numeric می‌باشد، می‌توان برای تبدیل نوع‌داده‌های numeric به nominal از فیلتر NumericToNominal بهره برد. داریم:

ID	Feature name	Feature type
1.	age	Numerical
2.	sex	Nominal
3.	chest pain type (4 values)	Nominal
4.	resting blood pressure	Numerical
5.	serum cholestoral in mg/dl	Numerical
6.	fasting blood sugar > 120 mg/dl	Nominal
7.	resting electrocardiographic results (values 0,1,2)	Nominal
8.	maximum heart rate achieved	Numerical
9.	exercise induced angina	Nominal
10.	oldpeak = ST depression induced by exercise relative to rest	Numerical
11.	the slope of the peak exercise ST segment	Numerical
12.	number of major vessels (0-3) colored by flourosopy	Numerical
13.	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect	Nominal
14.	Healthy or Not	Nominal

بخش دوم: بارگذاری و بررسی داده‌ها



حال اگر بخواهیم سه ویژگی که بهتر از بقیه می‌توانند برچسب کلاس را توصیف نمایند معرفی نمائیم، می‌توانیم از ابزار Select attributes نرم‌افزار weka استفاده نمائیم؛ که در آن می‌توان پس از تنظیم Attribute Evaluator به InfoGainAttributeEval و یا هم به GainRatioAttributeEval و نیز تنظیم Search Mode به Ranker، تمامی ویژگی‌ها را بنا به اندازه‌ی معیار مربوطه برای آن‌ها، به صورت نزولی مشاهده نمود. داریم:

Information Gain		GainRatio	
Attribute	**InfoGain Value**	**Attribute**	**GainRatio Value**
13 thal	**0.208556**	**13 thal**	**0.171204**
3 chestPain	**0.192202**	**12 numMajVess**	**0.17015**
12 numMajVess	**0.165916**	**10 oldPeak**	**0.148019**
9 exercIndcAng	0.129915	9 exercIndcAng	0.142054
8 maxHeartRateAchv	0.12028	8 maxHeartRateAchv	0.123102
10 oldPeak	0.119648	3 chestPain	0.111511
11 slpPeakExcerSTsgmnt	0.109917	11 slpPeakExcerSTsgmnt	0.110026
2 sex	0.066896	2 sex	0.073773
1 age	0.056726	1 age	0.056747
7 restElecRslt		7 restElecRslt	0.022886

	0.024152		
6 fastBldSgr	0.000193	6 fastBldSgr	0.000318
5 serumChlst	0	5 serumChlst	0
4 restBldPrss	0	4 restBldPrss	0

همانطور که پیداست سه ویژگی اول برای هر معیار، بهتر از سایرین می‌توانند برچسب کلاس را توصیف نمایند و همانطور که از نمودار پراکندگی کلاسی ویژگی‌ها نیز پیداست، می‌توان تخمین زد که مثلا برای معیار Information Gain ، سه ویژگی برتر قیدشده دارای پراکندگی کمتری در دسته‌بندی داده‌ها بوده و به عبارتی دارای آنتروپی کمتری می‌باشند و در نتیجه برای مسئله‌ی دسته‌بندی مناسب‌تر می‌باشند.

بخش سوم: ساخت درخت تصمیم

الف)

برای داده‌های آموزشی و البته انتخاب آخرین متغیر (HealthyOrNot) به عنوان متغیر هدف، confusion matrix به صورت زیر خواهد بود:

	A=1	B=2
A=1	145	5
B=2	9	111

معیارهای مختلف ارزیابی برای داده‌های آموزشی به قرار زیر است:

Class of interest	TP Rate (Recall)	FP Rate	Precision	F-Measure
1	.967	.075	.942	.954
2	.925	.033	.957	.941

ب)

Class of interest	F-Measure		
	β=.5	β=1	β=2
1	.947	.954	.962
2	.950	.941	.931

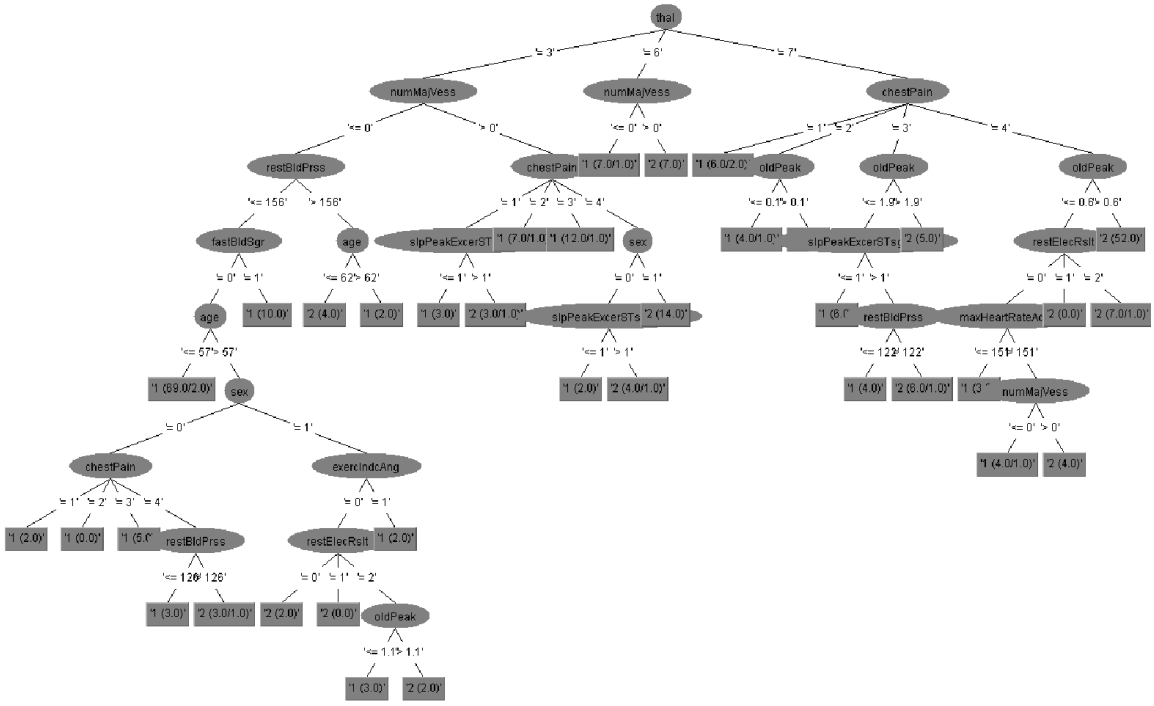
معیار F-Measure در واقع مقدار میانگین موزون (harmonic mean) حاصله از مقادیر Recall(Sensitivity) و Precision بوده و شدت سودمندی بازیابی اطلاعات را (با توجه این‌که یک کاربر، β بار به Recall بیشتر از Precision اهمیت می‌دهد) نشان می‌دهد.

بنابراین مثلا در این‌جا برای کلاس ۱، با توجه به اینکه مقدار خروجی weka برای Recall بیشتر از Precision می‌باشد، در نتیجه با افزایش پارامتر β ، مقدار F-Measure یا همان میانگین موزون(هارمونیک) بنا به حکم، به سمت

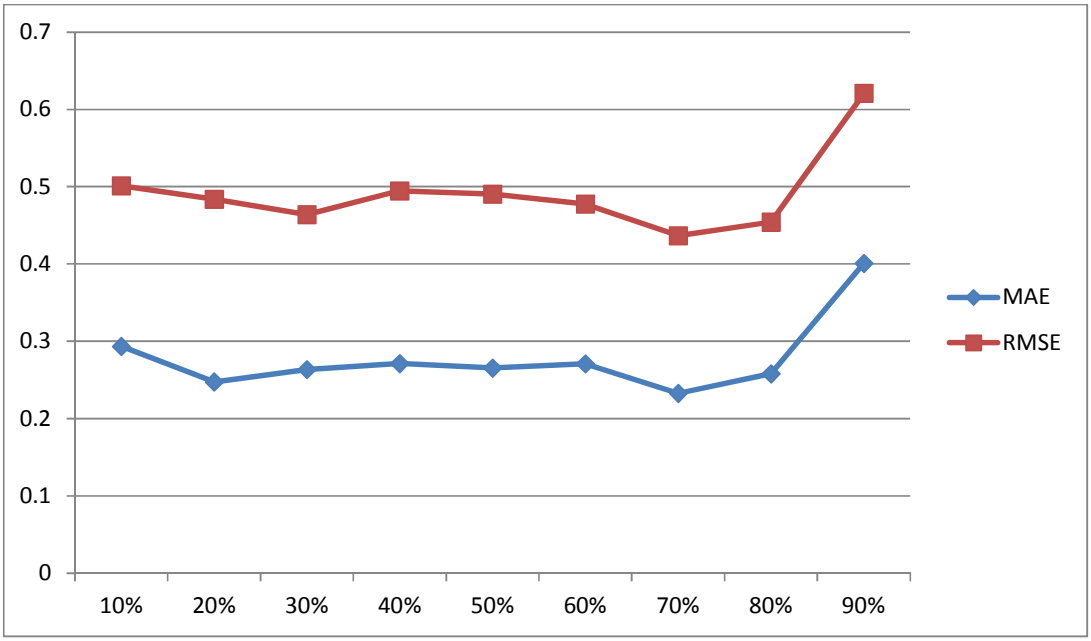
Recall که بیشتر می‌باشد متمایل گشته و افزایش می‌یابد و با کاهش β نیز کاهش می‌یابد. برای کلاس ۲ نیز با توجه به کمتر بودن مقدار Recall از Precision، با افزایش β مقدار F-Measure به سمت Recall که کمتر می‌باشد متمایل شده و کاهش می‌یابد و با کاهش β نیز از آن فاصله گرفته و افزایش می‌یابد.

لازم به ذکر است که به ازای مقدار یک برای پارامتر β ، نتایج حاصله از weka و قسمت «ب» با یکدیگر برابر می‌باشند و این نشان از آن دارد که نرم‌افزار weka در واقع از معیار $F_1 - Measure$ به جای $F_{.5} - Measure$ (که تأکید بیشتری بر Precision دارد تا Recall) و یا $F_2 - Measure$ (که تأکید بیشتری بر Recall دارد تا Precision) استفاده می‌نماید.

ج)



Training Data Percentage	Evaluation on test split				
	Correctly Classified Instances	Incorrectly Classified Instances	Total Number of Test Instances	Mean Absolute Error	Root Mean Squared Error
10 %	177 (72.84%)	66 (27.16%)	243	.2937	.5015
20 %	163 (75.46%)	53 (24.54%)	216	.2477	.4839
30 %	142 (75.13%)	47 (24.87%)	189	.2634	.4641
40 %	120 (74.07%)	42 (25.93%)	162	.2713	.4949
50 %	98 (72.59%)	37 (27.41%)	135	.2654	.4905
60 %	79 (73.15%)	29 (26.85%)	108	.2708	.4778
70 %	62 (76.54%)	19 (23.46%)	81	.2329	.4366
80 %	40 (74.07%)	14 (25.93%)	54	.2581	.4543
90 %	16 (59.26%)	11 (40.74%)	27	.4010	.6207



همانطور که پیداست به ازای اندازه‌ی مجموعه‌ی داده‌های آموزشی از ۱۰ تا ۸۰ درصد کل مجموعه‌ی داده، خطای تست (از هر دو نوع – MAE یا RMSE) چندان نوسانی نداشته و به ازای اندازه‌ی ۹۰ درصد، ناگهان میزان خطای تست افزایش

می‌یابد که این نشان از عمده‌ترین مشکل در الگوریتم‌های یادگیری یا همان OverFitting می‌باشد و به عبارتی Generalization الگوریتم ما پایین آمده و نسبت به داده‌های تست بد عمل می‌نماید.

ه)

Training Data Percentage	Tree Spec. (Unpruned)		Tree Spec. (Pruned)		
	Size of the Tree	Number of Leaves	Size of the Tree	Number of Leaves	Total Number of Test Instances
10 %	62	36	43	25	243
20 %	62	36	43	25	216
30 %	62	36	43	25	189
40 %	62	36	43	25	162
50 %	62	36	43	25	135
60 %	62	36	43	25	108
70 %	62	36	43	25	81
80 %	62	36	43	25	54
90 %	62	36	43	25	27

همانطور که پیداست در عین این‌که با عملیات هرس‌کردن، اندازه‌ی درخت حاصل کاهش یافته و این مسئله جهت جلوگیری از بیش‌برازش می‌باشد؛ اما با تغییر سهم داده‌های آموزشی اندازه‌ی درخت هرس‌شده و درخت هرس‌نشده تغییری نمی‌کند و این مسئله حاکی از آن است که الگوریتم ما بنا به هر تعداد از داده‌های آموزشی، درخت تصمیم یکسانی تولید می‌نماید.

و)

Training Data Percentage	Unpruned Tree	Pruned Tree
	RMSE	RMSE
60 %	.4778	.4666
100 %	.2035	.2430

همانطور که قابل مشاهده است، در حالت ۶۰ درصد داده‌ی آموزشی، با هرس‌کردن درخت تصمیم، میزان خطای RMSE کاهش می‌یابد؛ و این نشان از آن دارد که درخت نهائی با عملیات هرس‌کردن و حذف نودهای بی‌فایده، از حالت OverFitting خارج شده و نسبت به داده‌های تست عملکرد بهتری از خود نشان می‌دهد. اما در مورد ۱۰۰ درصد داده‌های آموزشی مشاهده می‌شود که خطای RMSE افزایش می‌یابد و این نشان‌دهنده‌ی آن است که درخت تصمیم نسبت به داده‌های آموزشی OverFit شده و کوچکترین جزئیات را نیز در نظر گرفته است و در نتیجه با عملیات هرس‌کردن و اعمال مجدد داده‌های آموزشی به عنوان داده‌های تست، شاهد افزایش خطای تست خواهیم بود.

چرا که در مورد انتخاب بخشی از مجموعه‌ی داده به عنوان داده‌های آموزشی و نه همه‌ی آن، اگر اندازه‌ی درخت بسیار بزرگ باشد، خطر OverFitting نسبت به داده‌های آموزشی و Generalization ضعیف نسبت به نمونه‌های جدید یا همان مجموعه‌ی تست افزایش می‌یابد و از طرفی یک درخت بسیار کوچک نیز ممکن است قادر به ضبط اطلاعات ساختاری

حائز اهمیت درباره‌ی فضای نمونه نباشد. هر چند دشوار است که بگوئیم در چه زمانی الگوریتم یادگیری درخت تصمیم باید توقف نماید، چرا که غیر ممکن است که بگوئیم فرضاً با اضافه‌کردن یک نود اضافه میزان خطا به سبک چشمگیری کاهش خواهد یافت و این مسئله البته تحت عنوان «اثر افق» شناخته می‌شود.

در اینجا نیز (۶۰ درصد داده‌های آموزشی) با هرس‌کردن درخت تصمیم نتایج تست بهتری حاصل شده و خطر بیش‌برازش نسبت به داده‌های آموزشی کاهش می‌یابد.

Test Option: 10-Fold Cross-validation

Unpruned							Pruned						
Tree Spec.		Classification Summary			Test Errors		Tree Spec.		Classification Summary			Test Errors	
Size of the Tree	Number of Leaves	Correctly Classified Instances	Incorrectly Classified Instances	Total Number of Instances	MAE	RMSE	Size of the Tree	Number of Leaves	Correctly Classified Instances	Incorrectly Classified Instances	Total Number of Instances	MAE	RMSE
85	50	199 (73.70%)	71 (26.30%)	270	.2779	.4854	58	33	197 (72.96%)	73 (27.4%)	270	.2856	.4747