سیستمهای خبرهٔ استقرایی و کاربرد آن در آمار

سید باقر میراشرفی* - حسن صادقی
دانشگاه فردوسی مشهد

خلاصه

سیستمهای خبره بر اساس هوش مصنوعی (AI) و با استفاده از زبانهای PROLOG , LISP و یا زبانهای برنامه سازی پیشرفته دیگر ساخته می‌شوند .

آمار و هوش مصنوعی بطور متقابل به توسعه یکدیگر کمک می‌کنند. از یک طرف روشهای آماری برای توصیف مفاهیم مبهم در سیستمهای خبره مورد استفاده قرار می‌گیرد. از طرف دیگر سیستم‌های خبره بعنوان ابزار فراگیری مفاهیم می‌توانند برای حل تعداد زیادی از مسائل طبقه بندی کردن آماری مورد استفاده قرار گیرد. بخصوص در روشهای تحلیل رگرسیون، تحلیل ممیزی، تحلیل خوشه ای و تحلیل سریهای زمانی کاربرد فراوانی دارد .

در این مقاله یکی از الگوریتمهایی را که کاربرد زیادی دارد مورد بررسی قرار می‌دهیم . این الگوریتم توسط Quinlan در سال ۱۹۷۹ تحت عنوان ID3 ، برای حالتی که متغیر کلاس کیفی باشد ، ارائه و تا سال ۱۹۸۶ اصلاح و تکمیل گردید. نتیجه این الگوریتم قواعدی است که بر اساس آن به مسائل طبقه بندی جواب داده می‌شود. در کنار این الگوریتم برای حالتی که متغیرها پیوسته باشند الگوریتم Regression Tree توسط Breiman (۱۹۸۴) مطرح می‌شود .

۱ـ مقدمه

تمامی سیستمهای خبره استقرائی بر اساس قاعده استقراء داده‌های آماری بوده و متعلق به دسته‌ای از روشها تحت عنوان "یادگیری از روی مثالها" می‌باشد. یادگیری از روی مثالها را می‌توان جزئی از تکنیک یادگیری ماشینی (ML) دانست. مضاف بر اینکه یادگیری ماشینی بخشی از هوش مصنوعی را تشکیل می‌دهد.

در کنار سیستم‌های خبره استقرائی، تعدادی از الگوریتم‌های شبکه عصبی و نیز روشهای آماری کلاسیک، تغییر کرسیون، سریهای زمانی وتحلیل ممیزی، از طریق یادگیری از روی مثالها تحلیل می‌شوند.

سیستم‌های خبره استقرائی پیشرفت اساسی در جهت تهیه فرآیندهای اتوماتیک و فراگیری دانسته‌ها داشته است. با استفاده از تجربیات گذشته که بصورت نمونه‌هائی در دسترس هستند، می‌توانیم قوانینی بر اساس دانسته‌ها تولید کنیم. بیشتر سیستم‌های خبره استقرائی که درحال حاضر مورد استفاده قرار می‌گیرد و کاربرد تجاری نیز پیدا کرده است، بر اساس مقاله مشترک Quinlan (۱۹۸۶) و Breiman (۱۹۸۴) می‌باشد.

که در بخش بعدی به معرفی این الگوریتمها و کاربرد آن می‌پردازیم.

۲ـ تحلیل ممیزی (طبقه بندی)

در تحلیل ممیزی (طبقه بندی) به روش کلاسیک، عموماً بر اساس روش کمترین مربعات خطا و مینیمم کردن تابع زیان، و با استفاده از نمونه‌ها، به توابع ممیزی مشخص جهت تحلیل دسترسی پیدا می‌کنیم.

اما در روشهای سیستم‌های خبره استقرائی بر اساس الگوریتم‌های مشخص، که توضیح یکی از این الگوریتمها بعداً خواهد آمد، به یک درخت تصمیم گیری رسیده و بر

اساس درخت، قواعدتصمیم‌گیری مشخص خواهدشد.وبا استفاده از قواعد بدست آمده، به تحلیل مسائل مورد نظر می‌پردازیم .

۳ـ الگوریتم ID3 (۱۹۸۶)

اولین بار درسال ۱۹۷۹ این الگوریتم توسط Quinlan مطرح که پس از اصلاح در سال ۱۹۸۶ بصورت کاملتری تحت عنوان "Intractive Dechotomizer 3" ارائه گردید.در این روش با استفاده از اطلاعات نمونه‌ها وبراساس دستورالعمل های زیر به نتایج مورد نظر خواهیم رسید.فرم کلی داده‌ها در جدول شماره ۱ نشان داده می‌شود.این الگوریتم زمانی مورد استفاده قرار می‌گیرد، که مقادیر مربوط به طبقه‌ها (کلاسها) مقادیر کیفی (گسسته) باشند .

دستور العمل ID3

۱ـ ابتدا کل نمونه‌هارا بعنوان یک مجموعه S در نظر می‌گیریم، که اصطلاحا" به آن گره نامیم .

۲ـ اگر تمام نمونه‌ها در گره S دارای کلاس مشترک C باشند، آنگاه دسته بندی مربوط به این گره C خواهد بود وبه دستورالعمل بعدی می‌پردازیم .

در غیر اینصورت خصیصه A راکه‌دارای بیشترین اطلاعات می‌باشد انتخاب می‌کنیم. سپس S را به مجموعه های جدا از هم بر اساس مقادیر A افراز می‌کنیم وهر یک از این مجموعه‌های افراز شده را بعنوان گره S در نظر گرفته ودستور العمل مرحله ۲ را برای تمام این گره ها اجرا می‌کنیم .

۳ـ در این مرحله تمامی گره‌های نهایی کلاس یکسانی داشته والگوریتم خاتمه پیدا می‌کند .

جدول شماره ۱ ـ فرم کلی داده‌ها در یادگیری از روی مثالها

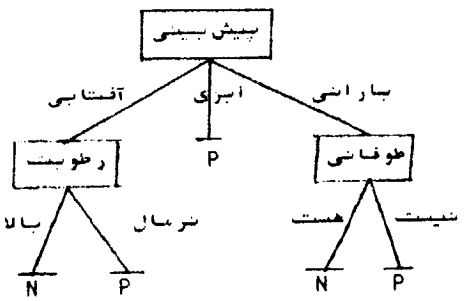
طبقه‌ها (کلاسها)	خصیصه‌ها (متغییرها)					نمونه‌ها
C_1	α_{11}	α_{12}	.	.	α_{1m}	۱
C_2	α_{21}	α_{22}	.	.	α_{2m}	۲
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
C_n	α_{n1}	α_{n2}	.	.	α_{nm}	n

مثال ۱ـ در یک تحقیق وضعیت هوای صبح چهارده روز مورد مطالعه قرار گرفت. و هدف از این تحقیق این است که بر اساس مقادیر متغیرهای مورد بررسی، از مثبت بودن شرایط هوا برای یک فعالیت به نوص، مطلع شویم. متغیرها عبارتند از پیش بینی هوا با مقادیر آفتابی، ابری و بارانی و متغیر دما با مقادیر سرد، معتدل و گرم، همچنین متغیر رطوبت با مقادیر بالا و نرمال، متغیر طوفانی با مقادیر درست و غلط. اطلاعات مربوط به این تحقیق در جدول شماره ۲ آمده است.

جدول شماره ۲ ـ یک مجموعه اجرائی کوچک

کلاس‌ها	خصیصه‌ها (متغییرها)				نمونه‌ها
	طوفانی	رطوبت	دما	پیش‌بینی‌هوا	
N	نیست	بالا	گرم	آفتابی	۱
N	هست	"	"	"	۲
P	نیست	"	"	ابری	۳
P	"	"	معتدل	بارانی	۴
P	"	نرمال	سرد	"	۵
N	هست	"	"	"	۶
P	"	"	"	ابری	۷
N	نیست	بالا	معتدل	آفتابی	۸
P	"	نرمال	سرد	"	۹
P	"	"	معتدل	بارانی	۱۰
P	هست	"	"	آفتابی	۱۱
P	"	بالا	"	ابری	۱۲
P	نیست	نرمال	گرم	"	۱۳
N	هست	بالا	معتدل	بارانی	۱۴

با استفاده از اطلاعات جدول فوق (مجموعه اجرائی) وبر اساس الگوریتم ID3، درخت تصمیم گیری نمودار ۱ حاصل می‌شود. که بر اساس درخت حاصل قواعد تصمیم گیری بصورت ذیل خواهد شد .



نمودار ۱ ـ درخت تصمیم گیری ساده

قواعد تصمیم گیری

۱ـ اگر هوای صبح روزشنبه آفتابی ورطوبت بالا باشد آنروز برای فعالیت بخصوص نامساعد است .

۲ـ اگر هوای صبح روزشنبه آفتابی ورطوبت نرمال باشد آنروز برای فعالیت بخصوص مساعد است .

۳ـ اگر هوای صبح روزشنبه ابری باشد آنگاه آنروز برای فعالیت بخصوص مساعد است .

۴ـ اگر هوای صبح روزشنبه بارانی وطوفانی هم باشد آنگاه آنروز برای فعالیت بخصوص نامساعد است .

۵ـ اگر هوای صبح روزشنبه بارانی وطوفانی نباشد آنگاه آنروز برای فعالیت بخصوص مساعد است .

در دستورالعملهای ID3 ، بعضی مواقع درمراحله‌دوم لازم‌است‌که خصیصه دارای بیشترین اطلاعات را تعیین کنیم .برای این منظور از تکنیک زیر استفاده می‌کنیم .

در این روش ابتدا اطلاعات مربوط به کل نمونه‌ها را بدست آورده وسپس بر اساس آن، میزان موثر بودن تمامی متغیرها را محاسبه می‌کنیم و درمیان این مقادیر ، بزرگترین مقدار را مشخص کرده ودر نتیجه متغیر مربوطه ، دارای بیشترین اطلاعات می‌باشد .

فرض کنید متغیر کلاس مربوط به مجموعه اجرائی وجدول دادها دارای K رده باشد .همچنین فرض کنید خصیصه (متغیر) A دارای مقادیر A1 و A2 و ... و Av و تعداد نمونه‌های مجموعه (کره)S برابر n باشد .برای محاسبه میزان عدم اطلاع غیر شرطی کلاس (آنتروپی کل کلاس) از فرمول زیر استفاده می‌کنیم .

$$I(C_1,\dots,C_K)=-\sum_{i=1}^{K}P_i\log_2 P_i \quad , \quad P_i=\frac{\#C_i}{n}=\frac{\text{تعداد نمونه‌های دارای کلاس } C_i}{\text{تعداد کل نمونه‌ها در } S}$$

که در آن

اکنون برای محاسبه میزان موثر بودن متغیر دلخواه A ، ابتدا متوسط آنتروپی کلاس با فرض متغیر را بدست آورده و از آنتروپی کل متغیر کلاس کم کرده و مقدار حاصل میزان موثر بودن A برای پیش بینی کلاس خواهدبود. وبرای بدست آوردن متوسط آنتروپی کلاس با فرض متغیر A ، براساس فرمول میانگین موزون زیر عمل می‌کنیم .

$$E(A)=\sum_{i=1}^{V}\frac{\#A_i}{n}I(C_1,C_2,\dots,C_K/A_i)$$

$$I(C_1,C_2,\dots,C_K/A_i)=-\sum_{i=1}^{K}P_i\log_2 P_i$$

که در آن

$$P_i=\frac{\#(C_i/A_i)}{\#A_i}=\frac{\text{تعداد نمونه‌هایی که براساس افراز } A_i \text{ دارای کلاس } C_i \text{ هستند}}{\text{تعداد نمونه‌هایی که در افراز } A_i \text{ قرار دارند}}$$

ودر آن

ودرنتیجه :

$$\text{میزان موثر بودن } A = I(C_1,C_2,\dots,C_K)-E(A)$$

این میزان برای تمامی متغیرها محاسبه وماکزیمم میزان را مشخص می‌کنیم که متغیر مربوطه ، همان متغیر دارای بیشترین اطلاعات خواهدبود. چون در اینجا I (C1,C2,...,Cn) برای تمامی متغیرها یکسان است. بنابر این کافی است. متوسط آنتروپی کلاس با فرض متغیرها رامحاسبه واز بین متوسط ها. می‌نیمم را مشخص کرده ومتغیر مربوطه همان متغیر دارای بیشترین اطلاعات خواهدبود .

این محاسبات‌را برای مثال ۱ انجام میدهیم. در اینجا تعداد نمونه‌ها ۱۴= n و تعداد کلاسها ۲= K وبعنوان مثال برای متغیر پیش بینی ۳= V است. ابتدا برای متغیر پیش بینی که دارای سه مقدار آفتابی، ابری وبارانی می‌باشد متوسط آنتروپی را بدست می‌آوریم .

برای مقدار آفتابی داریم : #N1 = ۳ و #P1 = ۲ وبنابراین

$$I(N,P/A_1 = \text{آفتابی}) = -\frac{۹}{۱۴}\log_۲\frac{۹}{۱۴} - \frac{۵}{۱۴}\log_۲\frac{۵}{۱۴} = ۰/۹۴ \text{ بایت}$$

وبطور مشابه برای مقادیر ابری وبارانی داریم :

$$I(N,P/A_۲ = \text{ابری}) = ۰ \quad \text{و} \quad \#P_۲ = ۴ \text{ و } \#N_۲ = ۰$$

$$I(N,P/A_۳ = \text{بارانی}) = ۰/۹۷۱ \quad \text{و} \quad \#P_۳ = ۳ \text{ و } \#N_۳ = ۲$$

ودر نتیجه :

$$E(\text{پیش بینی}) = \frac{۵}{۱۴} I(N,P/A_1 = \text{آفتابی}) + \frac{۴}{۱۴} I(N,P/A_۲ = \text{ابری}) + \frac{۵}{۱۴} I(N,P/A_۳ = \text{بارانی}) = ۰/۶۹۴$$

وبنابراین

بایت ۰/۲۴۶ = (پیش‌بینی) E − ۰/۹۴ = میزان موثر بودن متغیر پیش‌بینی

وبطور مشابه برای دیگر متغیرها داریم .

بایت ۰/۰۲۹ = موثر بودن دما

بایت ۰/۱۵۱ = موثر بودن رطوبت

بایت ۰/۰۴۸ = موثر بودن طوفانی

و از روی این مقادیر واضح است که متغیر پیش بینی بعنوان متغیر دارای بیشترین اطلاعات تعیین می‌شود .ونمونه‌ها بر اساس مقادیر این متغیر تقسیم می‌شوند وهر افرازی از این متغیر بعنوان یک گره (مجموعه S) در نظر گرفته وبا توجه به متغیرهای باقیمانده ، مرحله دوم دستورالعمل‌های ID3 را روی آن پیاده می‌کنیم .یکی از حالتهای خاص وقتی اتفاق می‌افتد که یکی از مجموعه‌های افراز شده (گره‌ها) خالی از نمونه باشد .در این حالت ID3 برای آن گره مقدارتهی (NULL) را در نظر گرفته ودرپایان اجرای الگوریتم ، به گره‌های تهی، یکی از کلاسهایی را که بیشترین فراوانی را در مجموعه نمونه‌ها داشته

باشد، انتخاب می‌دهد .

حالت خاص دیگروقتی اتفاق می‌افتد که یکی از متغیرها دقیقا " منطبق بر طبقه بندی باشد .فرض کنید متغیر A دارای مقادیر A1 و ...و Av و مجموعه نمونه‌های S را بر اساس مقادیر متغیر A افراز می‌کنیم .چنانچه برای هریک از مجموعه‌های افراز شده ، نسبت‌نمونه‌هایی را که دارای کلاس C (i=۱,۲,..., K) هستند .برابر نسبت نمونه‌ها در کل مجموعه S دارای آن کلاس باشند ، آنگاه این متغیر را دقیقا " منطبق برطبقه بندی گوئیم .بنابراین وجود این متغیر در درخت تصمیم گیری ضرورتی نخواهد داشت .

الگوریتم ID3 برای حل این مسئله بر اساس میزان موثر بودن متغیرها ، از متغیرهای غیرموثر صرف نظر می‌کند .ودرخت تصمیم گیری را بر اساس متغیرهای موثر تشکیل می‌دهد ، بدینصورت که چنانچه برای هر متغیر ، میزان موثر بودن آن از قدر مطلق یا درصدی از یک معیار مشخص کوچکتر باشد ، متغیر را غیر موثر شناخته و در محاسبات شرکت نمی‌دهد .در غیر اینصورت موثرشناخته وبراساس روشی که گفته شد ، بدنبال موثرترین متغیر وسپس ادامه الگوریتم می‌باشد .

این الگوریتم برای مقادیر گمشده واطلاعات دارای نوسان (NOISE) نیز راه حلهای مناسبی ارائه می‌دهد .جهت اطلاع بیشتر در این زمینه وهمچنین در رابطه با الگوریتم Regressio Tree می‌توانید به پایان نامه تحصیلی اینجانب سید باقر میراشرفی دانشکده علوم دانشگاه فردوسی مشهد رجوع نمائید .

نرم افزارهایی نیز در این رابطه از سوی شرکت های مختلف تهیه ومورد استفاده قرار می‌گیرند .

دو نمونه از این نرم افزارها تحت عناوین NEWID , KET وده که بهمراه راهنمای مربوطه در اختیار قرار دارد .درکنار این دو نرم‌افزار ، نرم‌افزار

دیگری نیز اینجانب با همکاری گروه کامپیوتر دانشگاه فردوسی مشهد تهیه نموده ومورد استفاده قرار دادیم .

۴ـ نتیجه گیری

همانگونه که در این مقاله آمده است، مابدنبال سیستمی هستیم که با استفاده از آن بتوانیم براحتی وبا دقت بیشتر ونیز سرعت زیاد، تجزیه وتحلیلهای مختلف آماری را که در زمینه‌های مختلف کاربردی مورد استفاده قرار میگیرد، به انجام برسانیم .

برای نیل به این هدف از سیستم‌های خبره ای که بر اساس روند فکر کردن افراد خبره تشکیل شده باشد، استفاده میکنیم.واانتظار داریم که در آینده ای نه چندان دور برای تمامی تجزیه وتحلیلهای آماری، سیستم‌های خبره ای تهیه و مورد استفاده علوم ودر خدمت بشر قرار گیرد .

در پایان لازم میدانم از جناب آقای ازقامی استادیار دانشگاه فردوسی مشهد به خاطر زحماتی که در رابطه با این مقاله متقبل شدند تشکر وقدردانی نمایم .

مراجع :

۱- پایان نامه کارشناسی ارشد آمار - سیدباقر میر اشرفی - دانشکده علوم - دانشگاه فردوسی " مشهد "

2) Breiman,L.et al(1984),Classification and Regression Tree.Wandsworth Belmont.

3) Nakhaeizadeh, G.(1991 a),Inductive Expert Systems and their Application in Statistics. Erscheint in Proceedings of Statsoft 91.

4) Nakhaeizadeh, G.(1991b),Application Of Machine Learning to solving industrial problems. Eyscheint in Proceedings of SOR 91 Trier, Physica Verlag.

5) Quinlan,J.R.(1986), Induction of Decision Trees.Machine Learning. VOL. 1,81-106.