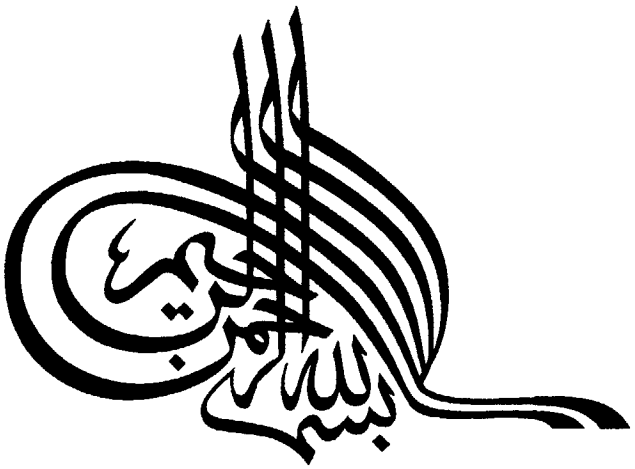
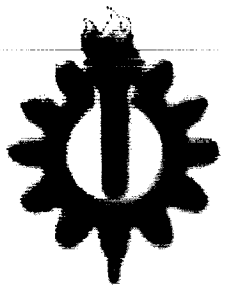


٣٩٩٨٤

۱۳۸۰ /۱۲/ ۲۸



مرکز اطلاعات و مدارک علمی ایران
تقدیم به مرکز

دانشگاه علم و صنعت ایران

دانشکده مهندسی کامپیوتر

016522

۳۹۹۸٤

یادگیری درخت تصمیم در محیط‌های مغشوش با بهره‌گیری از پردازش موازی

ستار هاشمی

پایان نامه‌ای برای دریافت درجه کارشناسی ارشد
در رشته مهندسی کامپیوتر
گرایش هوش مصنوعی

استاد راهنما : دکتر محمد رضا کنگاوری

پائیز ۱۳۸۰

۳۹۹۸٤

تقدیم

به پدر و مادر عزیزم

چکیده

درخت تصمیم یکی از روشهای نمایش دانش بوسیله یادگیرنده‌های استقرایی می‌باشد که طی بیست سال گذشته مورد توجه بوده است. استقراء درخت تصمیم در محیطهای مغشوش شامل دو مرحله توسعه درخت و هرس آن می‌باشد که هر کدام نقش به سزایی در افزایش دقت درخت تصمیم نهایی ایفا می‌کند. اخیرا استفاده از پردازش موازی به منظور توسعه درخت تصمیم مطرح شده است که زمان یادگیری بر روی مجموعه‌های آموزشی بزرگ را کاهش می‌دهد. در این پایان‌نامه یک روش جدید موازی‌سازی در سطح گره جهت توسعه درخت پیشنهاد شده است که با استفاده از ترکیب چندین روش یادگیری بصورت موازی سعی در بهبود دقت درخت حاصله دارد. در این روش ابتدا مجموعه آموزشی به صورت تصادفی به چند قسمت همپوشان تقسیم شده و هر قسمت در اختیار یک پردازنده قرار میگیرد. پردازنده‌ها پس از انجام عمل یادگیری در هر گره، نتایج را در اختیار پردازنده دیگری (ناظر) قرار میدهد، این پردازنده با استفاده از دو مدل فازی و کرامر، درخت نهایی را تولید می‌کند. در صورتیکه تعداد نمونه‌های مجموعه آموزشی کم نبوده و توزیع کلاسها و خصیصه‌ها در آن یکنواخت باشد، دقت درخت مورد نظر بهبود می‌یابد.

با تشکر

از دکتر کنگاوری به خاطر کمکهایی که در اجرای این پروژه نمودند

فهرست مطالب

عنوان **صفحه**

مقدمه ۱
فصل اول ۳
۱-۱) مقدمه ۴
۱-۲) ماشین یادگیرنده ۵
۱-۲-۱) اجزاء ماشین یادگیرنده ۶
۱-۲-۲) زبان توصیفی ۶
۱-۲-۳) استراتژی یادگیری ۷
۱-۲-۴) استقراء سازنده ۹
۱-۲-۵) استقراء سازنده با استراتژی چندگانه ۹
۱-۳) الگوریتم های یادگیری تولید کننده درخت ۱۱
۱-۳-۱) ID3 ..... ۱۱
۱-۳-۲) Assistant 86 ..... ۱۳
۱-۳-۳) کلاسبند اریب OC1 ..... ۱۴
۱-۴) الگوریتم های یادگیرنده تولید کننده قوانین ۱۷
۱-۴-۱) AQ15 ..... ۱۷
۱-۴-۲) CN2 ..... ۲۰
۱-۴-۳) DUCE ۲۲
۱-۵) شبکه های عصبی در مقایسه با درختهای تصمیم ۲۵
۱-۶) جمع بندی ۲۷
فصل دوم ۲۸
۲ـ۱) مقدمه ۲۹
۲-۲) درخت تصمیم کلاسبند ۲۹
۲-۳) کاربرد در دنیای واقعی ۳۲
۲-۴) مجموعه های آموزشی ۳۳
۲-۵) انتخاب نمونه های مجموعه آموزشی ۳۵
۲-۶) گسسته سازی داده ها ۳۶
۲ـ۷) استراتژیهای استقراء درخت تصمیم ۳۸
۲ـ۷ـ۱) جدول توافقی ۳۸
۲-۷-۲) معیارهای خوبی تقسیم ۴۰
۲-۸) ارزیابی درخت تصمیم ۴۲
۲-۹) تخمین دقت کلاسبندی ۴۸
۲-۱۰) ارزیابی عملی روشهای مختلف استقراء ۵۱
۲-۱۱) خلاصه و جمع بندی ۵۲
فصل سوم ۵۴

۳-۱) مقدمه ۵۵
۳-۲) هرس درخت تصمیم ۵۵
۳-۲-۱) هرس هزینه-پیچیدگی(CCP) ۵۶
۳-۲-۲) هرس مقدار بحرانی (CVP) ۵۹
۳-۲-۳) هرس کمترین خطا(MEP) ۶۰
۳-۲-۴) هرس خطای کاهش یافته(REP) ۶۱
۳-۲-۵) هرس خطای بدبین (PEP) ۶۲
۳-۲-۶) هرس مبتنی بر خطا (EBP) ۶۴
۳-۲-۷) مقایسه عملی روشهای مختلف هرس درخت تصمیم ۶۷
۳-۳) درخت تصمیم فازی ۶۹
۳-۳-۱) دامنه خصیصه‌ها ۷۰
۳-۳-۲) مجموعه های فازی ، منطق فازی و استنتاج تقریبی ۷۰
۳-۳-۳) الگوریتم فازی ID3 ، (FID3) ۷۴
۳-۳-۴) استنتاج درخت تصمیم فازی ۷۹
۳-۳-۵) افراز فضای خصیصه ارزش در درخت فازی ۸۰
۳-۳-۶) هیوریستیک Yuan برای ساخت درخت تصمیم فازی ۸۳
۳-۴) مقایسه درخت های تصمیم هرس شده با درخت تصمیم فازی ۸۴
۳-۵) جمع بندی ۸۵
فصل چهارم ۸۷
۴-۱) مقدمه ۸۸
۴-۲) یادگیری موازی درخت تصمیم ۸۹
۴-۳) انتخاب هیوریستیک برای یادگیری موازی ۹۳
۴-۳-۱) نگاه به جلو در درخت تصمیم ۹۴
۴-۳-۲) نگاهی دقیق تر به معیار کای اسکور ۹۸
۴-۴) موازی سازی در سطح گره درخت تصمیم ۱۰۱
۴-۵) بررسی یک مثال ۱۰۴
۴-۶) شبیه‌سازی و نتیجه آزمایشات بر روی مجموعه‌های آموزشی ۱۰۸
۴-۷) خلاصه و جمع‌بندی ۱۲۰
فصل پنجم ۱۲۲
۵-۱) خلاصه ۱۲۳
۵-۲) تحقیقات آتی ۱۲۶
مراجع ۱۳۲
واژه‌نامه ۱۳۶

فهرست اشکال

عنوان **صفحه**

شکل ۲ـ۱) یک درخت تصمیم (مجموعه آموزشی گلف) ۳۱

شکل ۲ـ۲) انتخاب نقطه قطع بین دو نقطه مرزی ۳۷

شکل ۳-۱) قسمتی از یک زیر درخت هرس شده ۵۶

شکل ۳-۲)کاهش در میزان خطا ۵۸

شکل۳-۳) درخت T' با هرس کردن درخت T در گره و ۱ بدست می آید در حالیکه T'' با
حرکت دادن زیر درخت دارای گره ریشه ۳ به مکان گره ۱ بدست می آید (grafting) ۶۵

شکل ۳-۴) تقسیم‌بندی مجموعه آموزشی برای توسعه و هرس درخت (الگوریتم‌هایی که نیاز به مجموعه هرس
جداگانه ندارند از اجتماع این دو مجموعه، هم برای رشد و هم هرس درخت استفاده میکنند) ۶۸

شکل۳-۵) زیرمجموعه‌های فازی برای متغیر فازی درآمد (Income) و مقدار عضویت برای ورودی خام u ۷۱

شکل۳-۶) استنتاج از قوانین فازی ۷۴

شکل ۳-۷) درخت تصمیم فازی ۷۵

شکل۳-۸) زیردرخت فازی تولید شده ۷۹

شکل۳-۹) استنتاج فازی در درخت تصمیم فازی ۸۰

شکل۳-۱۰) قطعه بندی (افراز) فضای خصیصه ارزش عددی ۸۲

شکل۴-۱) یادگیری موازی درخت تصمیم با استفاده از چهار پردازنده ۹۱

شکل۴-۲) درخت تولید شده بر روی مجموعه آموزشی Iris و قوانین مربوطه ۹۱

شکل۴-۳) اثر نگاه یک سطح به جلو بر روی عمق، دقت و اندازه درختن تولید شده
با بهره اطلاعاتی (اعداد منفی نشانگر بدتر شدن هر کدام از ویژگیها میباشد) ۹۵

شکل۴-۴) اثر نگاه یک سطح به جلو برروی دقت کلاسبندی تعدادی از مجموعه‌های
آموزشی (دقت‌های مربوط به نگاه به جلو با هرس مقایسه شده است ۹۶

شکل۴-۵)درخت تصمیم با معیار بهره اطلاعاتی ۹۸

شکل۴-۶) درخت تصمیم با نگاه یک سطح به جلو ۹۸

شکل۴-۷)یادگیری موازی در سطح گره ۱۰۱

شکل۴-۸) خصیصه‌ها و نمونه‌های مربوط به مجموعه آموزشی Golf ۱۱۱

شکل۴-۹) درخت ساخته شده بر روی مجموعه آموزشیGolf با استفاده از ضریب کرامر ۱۱۳

شکل۴-۱۰) میزان کاهش خطا در مجموعه‌های آموزشی با تعداد نمونه‌های مختلف (این اعداد نرمال شده‌اند) ۱۱۷

شکل۴- ۱۱) میزان کاهش خطا در مجموعه‌های آموزشی با مبنای خطای مختلف (این اعداد نرمال شده‌اند) ۱۱۸

شکل ۵-۱)یک مدل برای ساخت درخت تصمیم همکار ۱۲۸

فهرست جداول

عنوان **صفحه**

جدول۲-۱) مجموعه‌های آموزشی ۳۳
جدول۲-۲) ارزش دو خصیصه از دامنه سرطان سینه ۳۹
جدول۲ـ۳) جدول توافقی برای خصیصه اشعه ۳۹
جدول۲-۴) جدول توافقی برای خصیصه یائسگی ۴۰
جدول۲-۵) فرم عمومی جدول توافقی ۴۰
جدول۲ـ۶) احتمال آنکه آماره‌های χ^2 وG یک مقدار بزرگتر مساوی مقدار محاسبه
شده از جداول ۲-۳و۲-۴ داشته باشند ۴۳
جدول۲-۷) درصد خطای کلاسبندی معیارهای مختلف تولید درخت تصمیم برروی
چهار مجموعه آموزشی (درخت‌های نهایی هرس شده‌اند) ۵۱
جدول۳-۱) نتیجه اعمال روشهای مختلف هرس بر روی قدرت پیش‌بینی درخت
ساخته شده از چند مجموعه آموزشی ۶۸
جدول۴-۱) فراوانی نمونه‌ها با ارزشها و کلاسهای مختلف ۹۹
جدول۴-۲) میزان خطای روشهای مختلف تولید درخت تصمیم (٪) ۱۱۴
جدول۴-۳) میانگین میزان خطای روشهای مختلف تولید درخت تصمیم (٪) ۱۱۶

مقدمه

یک روش برای استخراج دانش از داده‌های با حجم زیاد استفاده از ماشین‌های یادگیرنده استقرائی می‌باشد. این ماشین‌ها دانش مربوط به مشاهدات، مستندات، گزارشات و ... را استخراج کرده و به یکی از دو شکل قوانین تصمیم و یا درخت‌های تصمیم که مورد بحث ما در این پایان‌نامه می‌باشد، فرموله می‌کنند. مشاهدات و نمونه‌های مورد یادگیری عموماً شامل تعدادی خصیصه-ارزش می‌باشند که این خصیصه‌ها دارای مقادیر گسسته و یا پیوسته هستند. هر نمونه دارای یک خصیصهٔ برچسپ کلاس میباشد که کلاس آن نمونه را مشخص می‌کند. درخت‌های تصمیم نیز متشکل از گره‌ها برای خصیصه آزمون، لبه‌ها که ارزشهای مختلف (گسسته) خصیصه آزمون را مشخص می‌کند و تعدادی برگ برای برچسپ کلاس‌ها می‌باشند. برای استنتاج درخت تصمیم کافی است با توجه به خصیصه آزمون در یک گره و نیز ارزش خصیصه متناظر آن در نمونه، یک مسیر را از ریشه درخت دنبال کرده تا به یک برگ برسیم، برچسپ آن برگ به عنوان کلاس نمونه برگردانده میشود. یادگیری درخت تصمیم که بوسیله کوینلن با پیشنهاد ID3 عمومیت پیدا کرد را بطور کلی می‌توان به دو مرحلهٔ انتخاب خصیصه غالب در گره جاری جهت توسعه درخت تصمیم و هرس درخت نهایی به منظور افزایش جامعیت و دقت درخت (بویژه در محیطهای مغشوش) تقسیم کرد. بدیهی است که انتخاب روش مناسب جهت توسعه درخت نقش کلیدی بر روی جامعیت درخت نهایی ایفا می‌کند. اگرچه روشهای مختلفی برای توسعه درخت تصمیم ارائه شده است اما نمی‌توان روشی را پیدا کرد که به طور مطلق در همه محیطها بهترین باشد، به همین دلیل استفاده از پردازش موازی در توسعه درخت تصمیم به منظور بهره‌گیری از ویژگی روشهای مختلف یادگیری، منطقی بنظر می‌رسد. یادگیری موازی را میتوان از دو نقطه‌نظر بررسی کرد؛ ۱) استفاده از پردازش موازی به منظور کاهش زمان یادگیری بر روی مجموعه‌های آموزشی بزرگ ، که مجموعه آموزشی در بین چندین پردازنده تقسیم شده و با توجه به ماهیت موازی ساخت درخت تصمیم، کار تولید درخت بوسیله پردازنده‌ها بصورت همزمان صورت میگیرد. در این روش همه پردازنده‌ها الگوریتم یادگیری یکسانی را بر روی مجموعه آموزشی اجرا می‌کنند. ۲) استفاده از یادگیرنده‌های مختلف به صورت موازی برای افزایش دقت یادگیری، که با هیوریستیک‌های مختلف چندین درخت بر روی مجموعه آموزشی ساخته میشود، سپس از همه این درخت‌ها برای کلاسبندی نمونه آینده استفاده میشود، در صورتیکه کلاسهای مختلفی بوسیله این درخت‌ها پیشنهاد شود، کلاسی که بیشترین فراوانی را در بین کلاسهای پیشنهادی بوسیله درختها دارد، بعنوان کلاس این

نمونه انتخاب میشود. اما روش پیشنهادی ما موازی‌سازی را در سطح گره‌ها انجام داده و در نهایت منجر به تولید تنها یک درخت خواهد شد. روش کار به این شکل است که ابتدا مجموعه آموزشی به چند زیرمجموعه همپوشان (در آزمایشات ما این تعداد معادل سه است) تقسیم شده و بصورت موازی بر روی هر کدام یک پردازنده عمل یادگیری را انجام می‌دهد. در هر گره، از بین خصیصه‌های پیشنهاد شده بوسیله پردازنده‌ها، با استفاده از دو مدل فازی و کرامر، بهترین خصیصه برای توسعه درخت انتخاب شده و درخت توسعه می‌یابد. مدل کرامر مبتنی بر رابطه ریاضی کرامر پیشنهاد شده است که قادر است روابط موجود بین خصیصه‌ها را بدست آورد، اما بدلیل متقارن بودن رابطه، این مدل نمی‌تواند از بین دو خصیصه یکی را انتخاب کند. این مدل در چنین شرایطی از رأی‌گیری در سطح نود استفاده می‌کند. مدل فازی نیز بر اساس یک رابطه فازی پیشنهاد شده است که بدلیل نامتقارن بودن قادر است از بین دو خصیصه یکی را انتخاب کند. نتایج نشان می‌دهد که در صورتی مجموعه آموزشی کوچک نباشد و همچنین در مجموعه آموزشی کلاسها و خصیصه‌ها دارای توزیع یکنواخت باشند، دقت درخت نهایی بهبود می‌یابد.

فصل اول این پایان‌نامه به بررسی ماشین‌های یادگیرنده پرداخته و چند مدل مختلف آن را بررسی می‌کند، فصل دوم با معرفی دقیق‌تر درخت تصمیم روشهای توسعه درخت را معرفی می‌کند، فصل سوم دو روش افزایش دقت و کاهش پیچیدگی در درخت تصمیم، یعنی هرس و منطق فازی را معرفی کرده و آنها را اجمالا مقایسه میکند، فصل چهارم با اشاره‌ای گذرا به روشهای موازی یادگیری درخت تصمیم، روشی جدید برای موازی سازی یادگیری درخت تصمیم ارائه کرده و با ارائه نتایج آن بر روی مجموعه‌های آموزشی این روش را ارزیابی می‌کند و فصل پنجم جمع‌بندی و تحقیقات آتی میباشد.

فصل اول

ماشین‌های یاد گیرنده

دانشگاه صنعتی اراک

۱-۱) مقدمه

نظر به فن آوری جدید اطلاعات که سالانه باعث تولید تعداد زیادی کامپیوترهای قدرتمند تر و جدید تر میشود امروزه امکان جمع آوری ، انتقال ، ترکیب و ذخیره حجم زیادی از اطلاعات را با هزینه زیاد، عملی ساخته است. افزایش مستندات و تصاویر ، صداها و جداول و ... ما را به سمت استخراج اطلاعات از این داده ها رهنمون می سازد و با یک نگاه دقیق به داده ها درمی یابیم که داده ها به تنهایی اگر چه زیاد هم باشند کافی نیستند.

یک روش برای استخراج دانش از داده های با حجم زیاد بنام داده کاوی[1] می باشد که ماشین های یادگیرنده[2] یکی از کاربردهای آن است. این ماشین ها دانش مربوط به نمونه ها و مشاهدات و یا مستندات که داده های ما هستند را استخراج می کنند . برای درک بهتر اختلاف بین داده[3] و دانش[4] (اطلاعات) تعدادی از ویژگی های آنها را ذکر می کنیم.

داده :

- به یک نمونه بر می گردد. (یک نمونه از اشیاء ، افراد ، حوادث ، ...)
- ویژگی ها را جداگانه توصیف می کند.
- با حجم زیاد در دسترس است (آرشیو و بانک اطلاعاتی)
- امکان پیش بینی را به ما نمی دهد.

دانش:

- به کلاسهای نمونه بر می گردد (مجموعه ای از اشیاء ، افراد ، حوادث و ...)
- بصورت عمومی ساختار ، الگوها و ... را توصیف می کند.
- غالبا" اخذ آن مشکل است.
- قابلیت پیش بینی دارد.

---

[1]Data Mining

[2]Machine learning

[3]Data

[4]Information

همانطوریکه دیده میشود، دانش بسیار با ارزش تر از داده می باشد که ضرورت استفاده از ماشین های یاد گیرنده را بعنوان یکی از ابزارهای داده کاوی تأئید می کند.

۱-۲) ماشین یادگیرنده

امروزه ماشین های یادگیرنده توسط محققین هوش مصنوعی بسیار مورد توجه است. ماشین یادگیرنده در سیستم های هوشمند برای افزایش دانش وتغییر آن، افزایش کارایی و تصحیح اتوماتیک خطا استفاده میشود. یکی از کاربردهای ماشین های یادگیرنده که بیشتر مورد توجه قرار گرفته است اخذ دانش است. به این معنی که عمل یادگیری اطلاعات پایه را از محیط استخراج کرده و برای تحلیل حوادث آینده از آن بهره می گیرد.

ماشین های یادگیرنده را بطور کلی میتوان به دو دسته تقسیم کرد که عبارتند از یادگیری نمادی و یادگیری عددی[5] . شبکه های عصبی یک مدل از یادگیرنده های عددی میباشد و سعی در مدل کردن مغز انسان دارد. از کاربردهای شبکه های عصبی می توان به بازشناسی گفتار، تشخیص الگو ، سیستم های کنترلی وقفی و ... اشاره کرد . از سیستم های با رویکرد نمادی میتوان سیستم های AQ15 , CN2 , Duce , Clips را برشمرد که دانش را به زبان درخت تصمیم و یا قوانین تصمیم فرموله میکنند.

- یادگیری عددی بوسیله شبکه عصبی بصورت موازی ، افزایشی و قدرتمند است که بصورت تعدادی از گره‌ها که بوسیله تعدادی لبه به هم وصل میشود در نظر گرفته میشود. به هر کدام از این لبه ها یک وزن اختصاص داده میشود که قابل تغییر است (در طول مرحله آموزش) و هر گره دارای یک تابع فعالیت می باشد که با جمع فضایی ورودیها و اعمال یک تابع خاص مثل آستانه ممکن است هر گره فعال و یا غیر فعال باشد. فیچرهای هر نمونه بعنوان ورودی به شبکه داده میشود و لبه ها فعالیت های گره‌های قبل را با اعمال یک ضریب وزن به گره‌های بعدی انتقال میدهد . عملکرد این گره‌ها کاملاً موازی صورت می گیرد و در نهایت کلاس نمونه اینده را پیش بینی می کند. مرحله آموزش شبکه ممکن است با ناظر و بدون ناظر صورت بگیرد. در یادگیری با ناظر کلاس هر نمونه در مجموعه آموزشی مشخص می باشد و شبکه سعی خواهد کرد که تابع خطا را کمینه کند اما در آموزش بدون ناظر تابع خطایی نخواهیم داشت و شبکه الگوهای نمونه های ورودی را استخراج می کند.
- یادگیرنده های نمادی ، یادگیری را با جستجو در فضای جستجوی نمادی مشاهدات مجموعه آموزشی انجام داده و نتیجه بصورت ، عبارات منطقی ، قوانین ، درخت تصمیم و یا شبکه معنی ذخیره

می شود سیستم های با این رویکرد را میتوان بسته به استفاده از دانش زمینه[1] به دو دسته تقسیم کرد. گروه اول شامل ماشینهایی است که نیاز به دانش دامنه خاصی ندارند مثل ID3 , CN2 , Duce و نسبت به بقیه عمومیت بیشتری دارند و گروه دوم یادگیری را با حضور دانش زمینه ، مثل واقعیت ها هیوریستیک ها و بعضی گشتارهای[2] قوانین مختص به آن دامنه انجام میدهد این برنامه ها تک منظوره و کاملا وابسته به محیط یادگیری هستند که EBL[3] , AQ در این دسته قرار می گیرند.

۱-۲-۱) اجزاء ماشین یادگیرنده

ماشین یادگیرنده با محیط برای اخذ ورودی و ارائه دانش به کاربر (معلم ، انسان خبره یا اپراتور) ارتباط برقرار می کند به همین منظور نیاز به یک زبان توصیفی دارد بصورتی که داده ورودی و دانش خروجی به آن زبان فرموله شود. گویایی زبان توصیف برای توفیق و کارایی سیستم های یادگیرنده نقش حیاتی دارد.

دومین جزء ماشین‌های یادگیرنده یعنی اطلاعات ورودی به سیستم را میتوان به دو قسمت تقسیم کرد، اول آن دسته از نمونه ها که بوسیله آن پایگاه دانش شکل می گیرد و دیگری نمونه هایی که پایگاه دانش حاصله بوسیله آن ارزیابی شده و درصورت نیاز پایگاه دانش دستکاری میشود. دریافت، پردازش و پرداخت اطلاعات بعنوان استراتژی یادگیری شناخته میشود که از مهمترین اجزاء یادگیری است و مورد توجه بسیاری از محققین دراین شاخه قرار گرفته است.

۱-۲-۲) زبان توصیفی[1]

همانطوریکه قبلا گفته شده زبان توصیف برای دریافت ورودی ، ارائه دانش خروجی و نمایش پایگاه داده استفاده می شود. قوانین ، شبکه های معنی و قابها، نمونه های مختلف زبان توصیف می باشند که بوسیله یادگیرنده های مختلف استفاده شده اند. این باید بتواند توصیف جامعی از دامنه مورد یادگیری را نمایش دهد چراکه نقش عمده ای را در یادگیری دارد. خروجی ماشین که به این زبان فرموله شده است بایستی قابل فهم باشد. زبان توصیف منطق بولی از کم هزینه ترین وکم دقت ترین روشهای نمایش دانش است، درحالیکه استفاده از قابها ، هزینه بالا، دقت بیشتر و قابلیت فهم بیشتر داشته و نیز قابلیت فرموله کردن دانش زمینه را دارد.

---

[1] Background Knowledge

[2] Transformator

[3] Expalanation Base Learning