دانشکده مهندسی

پایان‌نامه کارشناسی‌ارشد در رشته مهندسی کامپیوتر (هوش مصنوعی)

بهبود دقت طبقه بندی با استفاده از روش‌های انتخاب خصیصه‌ی ترکیبی

توسط:
فاطمه علیمردانی

استاد راهنما:
دکتر رضا بوستانی

شهریور ۱۳۸۸

به نام خدا

اظهارنامه

اینجانب فاطمه علیرضایی (۸۵۰۹۸۷) دانشجوی رشته‌ی مهندسی کامپیوتر گرایش هوش مصنوعی دانشکده‌ی برق و کامپیوتر

اظهارمی‌کنم که این پایان نامه حاصل پژوهش خودم بوده و در جاهایی که از منابع دیگران استفاده کرده‌ام، نشانی دقیق و مشخصات کامل آن را نوشته‌ام. همچنین اظهارمی‌کنم که تحقیق و موضوع پایان نامه‌ام تکراری نیست و تعهد می‌نمایم که بدون مجوز دانشگاه دستاوردهای آن را منتشر ننموده و یا در اختیار غیر قرار ندهم. کلیه حقوق این اثر مطابق با آیین‌نامه مالکیت فکری و معنوی متعلق به دانشگاه شیراز است.

نام ونام خانوادگی: فاطمه علیرضایی

تاریخ و امضا: ۸۸/۶/۱۸

به نام خداوند جان و خرد

به نام خدا

بهبود دقت طبقه بندی با استفاده از روش‌های انتخاب خصیصه‌ی ترکیبی

به وسیله‌ی :

فاطمه علیمردانی

پایان نامه
ارائه شده به معاونت تحصیلات تکمیلی دانشگاه به عنوان بخشی از
فعالیت‌های تحصیلی لازم برای اخذ درجه کارشناسی ارشد

در رشته‌ی:

مهندسی کامپیوتر - گرایش هوش مصنوعی

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی شده توسط کمیته پایان نامه با درجه: عالی
دکتر رضا بوستانی، استادیار بخش مهندسی کامپیوتر(رئیس کمیته).....
دکتر منصور ذوالقدری جهرمی دانشیار بخش مهندسی کامپیوتر.....
دکتر کریم رستگار استادیار دانشکده پزشکی دانشگاه علوم پزشکی شیراز.....
دکتر احمد غنی زاده، دانشیار دانشکده پزشکی دانشگاه علوم پزشکی شیراز.....

شهریور ۱۳۸۸

دکتر احمد غنی زاده
متخصص روانپزشکی، مرکز گسالان
فوق تخصص اعصاب و روان کودکان و نوجوانان
دانشیار تمام وقت دانشگاه علوم پزشکی شیراز ۰۷۱-۲۴۸۲۳۳

به:

آنهایی که به من آموختند حتی یک کلمه.

سپاسگزاری

در پایان این مسیر پرفراز و نشیب، که کوله‌باری از تجربه‌های گران‌بها را برایم به ارمغان آورده است، به طور خاص از استاد راهنمای محترم آقای دکتر رضا بوستانی، به خاطر صبوری‌ها و راهنمایی‌های ارزنده‌شان سپاسگزارم. همچنین از اساتید بزرگوارم جناب آقای دکتر کریم رستگار، جناب آقای دکتراحمد غنی زاده و آقای دکتر منصور ذوالقدر جهرمی که در طول این مدت از حمایت‌های ایشان برخوردار بوده‌ام کمال تشکر را دارم.

از تمامی دوستان به خصوص خانم خدیجه سادات نژاد و کلیه کادر بخش مهندسی و علوم کامپیوتر دانشگاه شیراز که با اینجانب همکاری صادقانه داشتند، تشکر می‌کنم و یک سپاس ویژه از بیماران محترمی که در انجام آزمایشات این تحقیق همکاری بی‌دریغی داشته‌اند دارم و برای تک تک آنها سلامت و شادی روز افزون آرزومندم.

چکیده

بهبود دقت طبقه بندی با استفاده از روش‌های نوین ترکیبی انتخاب خصیصه

به وسیله‌ی:
فاطمه علیمردانی

در سال‌های اخیر، کاراتر سازی روش‌های انتخاب خصیصه به صورت رو به رشد مورد توجه بوده‌اند. تحقیقات انجام شده به منظور کاهش هزینه‌ی محاسبات و همچنین کاهش ریسک تطابق بیش از حد مطالعه شده‌اند. در راستای کاهش بعد خصیصه‌ها از بین همه‌ی الگوریتم‌ها روش‌های پیشرونده به دلیل هزینه‌ی محاسباتی کم بسیار کارا هستند. در این رساله دو الگوریتم پیشرونده‌ی جدید بر اساس مقدار افزایش اطلاعاتی که با حضور هر خصیصه خواهیم داشت ارائه شده است. این الگوریتم‌ها سعی دارند خصیصه‌های حاوی بیشترین اطلاعات را یافته و انتخاب کنند. از طرف دیگر با غلبه بر محدودیت‌های الگوریتم‌های ترتیبی پیشرونده و پسرونده سعی می‌کنند زیرمجموعه‌ای را انتخاب کنند که خصیصه‌های آن بیشترین استقلال را داشته باشند. برای بررسی میزان خوبی خصیصه‌هایی که انتخاب می‌شوند از طبقه بندی کننده‌ی نزدیکترین همسایه استفاده شده است. این طبقه بند وابسته به معیار فاصله‌ای که برای یافتن نزدیکترین همسایه به کار می‌برد می‌باشد. در همین راستا یک نسخه‌ی تصحیحی برای حذف این وابستگی پیشنهاد شده است که از رای گیری بین سه فاصله نتیجه‌ی خود را برای نمونه‌ی آزمایشی اعلام می‌کند. نتایج، بهبود قابل توجه صحت طبقه بندی نسبت به روش‌های شناخته شده راروی داده‌های UCI نشان می‌دهند. دلیل این بهبود، کارایی الگوریتم‌های جدید معرفی شده است و از سویی دیگر طبقه بند پیشنهادی تصمیم پایدارتر و ثابت تری در انتخاب کلاس نمونه‌ی آموزشی نسبت به نسخه‌ی استاندارد آن می‌گیرد.

در یک آزمایش دیگر روش‌های پیشنهادی در تشخیص دو دسته بیماری روانی (دو قطبی و اسکیزوفرنی) که در پزشکی تشخیص کمّی روی آن‌ها وجود ندارد به کار برده شده است.این مطالعه آماری روی این دو بیماری برای اولین بار ارائه شده است.

فهرست مطالب

عنوان صفحه

۱- مقدمه ۱

۱-۱- انتخاب خصیصه ۱

۱-۲- ارزیابی یک زیر مجموعه از خصیصه‌ها ۳

۱-۲-۱- معیارهای مستقل ۳

۱-۲-۲- معیارهای وابستگی ۴

۲- مروری بر روشهای استخراج خصیصه از سیگنال ۹

۲-۱- سیگنالهای EEG ۱۱

۲-۲- تست پاسخ محرک بینایی ۱۲

۲-۳- ثبت سیگنال ۱۲

۲-۴- روشهای استخراج خصیصه ۱۸

۲-۴-۱- مدل خودبازگشتی ۱۹

۲-۴-۲- کورتوسیس ۲۱

۲-۴-۳- بعد فراکتالی و هندسه فراکتالی ۲۲

۲-۴-۳-۱-مولفههای لیاپانوف ۲۳

۲-۵- فیلترینگ ۲۷

۳- مروری بر روشهای گسسته‌سازی ۲۸

۳-۱- مقدمه ۲۸

۳-۲- معرفی گسسته سازی ۲۸

۳-۳- روشهای موجود ۲۹

۳-۴- ارزیابی نتایج گسسته‌سازی ۳۱

۳-۵- چهارچوب گسسته‌سازی ۳۲

۳-۶- روشهای تقسیم کردن ۳۳

۳-۶-۱- بخش‌بندی ۳۴

۳-۶-۲- معیار انتروپی ۳۷

۳-۶-۳- معیار وابستگی ۴۰
۳-۶-۴- معیار دقت ۴۱
۳-۷- روشهای ادغام کردن ۴۲
۳-۷-۱- معیار مربع کای ۴۲
۳-۷-۲- روش Chi Merge ۴۳
۳-۷-۳- روش Chi^2 ۴۳
۳-۸- تکنیک گسسته‌سازی بر اساس معیار فیشر ۴۵
۳-۸-۱- معرفی معیار فیشر ۴۵
۳-۸-۲- الگوریتم گسسته سازی بر اساس فیشر ۴۶
۴- مروری بر روشهای انتخاب خصیصه ۴۸
۴-۱- مقدمه ۴۸
۴-۲- الگوریتم جستجوی تبو ۴۹
۴-۳- پلاس. ال. ماینس. آر ۵۳
۴-۴- اف. سی. بی. اف ۵۳
۴-۴-۱- ارتباط خطی ۵۴
۴-۴-۲- اطلاعات متقابل ۵۵
۴-۵- شاخص دیویس بولدین ۵۹
۴-۶- رتبه بندی متغیرها ۶۳
۴-۶-۱- اصول متد و علامت گذاری ها ۶۳
۴-۶-۲- شرایط همبستگی ۶۳
۴-۷- مثالهای کوچک اما راه گشا ۶۴
۴-۷-۱- آیا متغیرهای اضافی فرضی به همدیگر کمک میکنند؟ ۶۵
۴-۷-۲- چگونه همبستگی بر افزونگی متغیرها تاثیر می گذارد؟ ۶۶
۴-۷-۳-آیا متغیری که به تنهایی مفید نیست در کنار بقیه متغیرها سودمند است؟ ۶۷
۴-۸- روشهای انتخاب خصیصه‌ی توکار(تو درتو) ۶۹
۴-۸-۱- متدهای زیرمجموعه های تودرتو ۶۹
۴-۸-۲- بهینه سازی اهداف مستقیم ۷۰
۴-۹- مقولات پیشرفته و مسایل باز ۷۱
۴-۹-۱- واریانس زیرمجموعه‌های انتخاب شده ۷۱
۴-۹-۲- معیارانتخاب خصیصه در مسائل بدون ناظر ۷۱
۴-۹-۳- انتخاب پیشرو در مقابل انتخاب عقبگرد ۷۲
۴-۱۰- مسایل چندکلاسی ۷۳

۵- طبقه بندی کننده ۷۴
۵-۱- مقدمه ۷۴
۵-۲- طبقه بند بیز ساده ۷۵
۵-۳- طبقه بند نزدیکترین همسایه ۷۶
۶- روشهای پیشنهادی و نتایج تجربی ۸۱
۶-۱- مقدمه ۸۱
۶-۲- روشهای پیشنهادی ۸۱
۶-۲-۱- الگوریتم حریصانه بر اساس ارتباط کلی خصیصه‌ها ۸۱
۶-۲-۲- الگوریتم سریع کاهش ویژگی با استفاده از شاخص دیویس بولدین ۸۶
۶-۳- مجموعه داده‌های سیگنال ۸۸
۶-۳-۱- روش شکاف آماری (گپ استاتیک) ۹۰
۶-۳-۲- روش ارزیابی Silhouette ۹۱
۶-۳-۳- خوشه بندی تفاضلی ۹۴
۶-۴- دادههای استاندارد ۹۹
7- نتیجه‌گیری و کارهای آینده ۱۰۴
۷-۱- نتیجه‌گیری ۱۰۴
۷-۲- کارهای آینده ۱۰۴
۸- مراجع ۱۰۶

فهرست جداول

عنوان صفحه

جدول 1- نقاط گسستگی داده‌ی با روش پهنای مساوی ۳۵
جدول 2- نقاط گسستگی داده‌ی Iris با روش فراوانی مساوی ۳۵
جدول ۳- نقاط گسستگی در روش 1R ۳۶
جدول 4- مقایسه سه روش بخش بندی ۳۷
جدول 5- مقایسه نتایج سه روش با آزمون مربع کای ۴۴
جدول 6 - تعداد خصیصه‌های انتخاب شده توسط روشهای مختلف. ۹۰
جدول ۷- دقت نتیجه شده از NN استاندارد در مقا یسه با NNپیشنهادی ۹۵
جدول ۸- دقت طبقه بند پیشنهادی بر روی هر کدام از خصیصه‌ها به تنهایی ۹۷
جدول ۹- نتایج انتخاب خصیصه +NN استاندارد و NN معرفی شده روی داده‌های نویزی ۹۷
جدول ۱۰- - مشخصات مجموعه داده‌های UCI به کار رفته در این رساله ۱۰۰
جدول ۱۱- - نتایج طبقه بندی کننده‌ی پیشنهادی روی خصیصه‌های انتخاب شده مختلف۱۰۱
جدول ۱۲- - نتایج طبقه بندی کننده‌های مختلف روی خصیصه‌های انتخابی GOR ۱۰۳

فهرست شکل‌ها

عنوان صفحه

شکل ۱ -دسته بندی روشهای انتخاب خصیصه[1] ۵
شکل ۲- مکان الکترودها در سیستم۲۰-۱۰ ۱۳
شکل ۳- الف) نمونهای از سیگنالهای ضبط شده الف) دوقطبی، ب) اسکیزوفرنی ۱۵
شکل ۴- طیف توان نمونه‌های از هر دو نوع بیماری به کمک Color Map، الف) دو قطبی، ب) اسکیزوفرنی ۱۷
شکل ۵- مدلسازی AR [77] ۲۰
شکل ۶- یک نمونه سیستم خطی[76] ۲۰
شکل ۷- الف)یک عملگر ساده فرکتال و اثر متوالی آن بر یک پاره خط. ب)یک شکل-فرکتال مصنوعی [80]. ۲۲
شکل ۸- منحنی جمع همبستگی برحسب شعاع کره برای یک سیگنال غیر نویزی. ۲۶
شکل ۹- چهارچوب سلسله مراتبی گسسته سازی[25] ۳۳
شکل ۱۰- مثال عملی مراحل انجام روش 1R ۳۶
شکل ۱۱- فلوچارت تکنیک گسسته سازی با معیار فیشر[74] ۴۷
شکل ۱۲- چهار گام کلیدی در فرایند انتخاب خصیصه[21] ۴۸
شکل ۱۳- شمای کلی الگوریتم تبو[15] ۵۱
شکل ۱۴- شبه کد الگوریتم تبو[20] ۵۲
شکل ۱۵- شبه کد الگوریتم پلاس. ال. ماینس. آر[83] ۵۴
شکل16 - شبه کدالگوریتم اف. سی. بی. اف[10] ۵۹
شکل ۱۷- کارایی روشهای مختلف انتخاب خصیصه تحت شرایط خاص[5] ۶۱
شکل ۱۸- اطلاعات به دست آمده از متغیرهای اضافی فرضی ۶۵
شکل ۱۹- تاثیر همبستگی بر افزونگی داده‌ها ۶۶
شکل ۲۰- بررسی مفید بودن متغیرها در همکاری با هم. ۶۹
شکل ۲۱- بررسی انتخاب عقبگرد و پیشرو ۷۳
شکل ۲۲- مثالی از عملکردNN با k=5 ۷۷
شکل ۲۳- کانتورهایی از فاصله‌های مساوی[28] ۷۹
شکل ۲۴- الگوریتم حریصانه بر اساس ارتباط کلی خصیصه‌ها. ۸۶

شکل ۲۵- شبه کد الگوریتم DB-FFR ۸۸
شکل ۲۶- مقدار Silhouett برای هر دو دسته بیماری ۹۳
شکل ۲۷- مقایسه‌ی نرخ تشخیص بین دو بیماری با الگوریتم های مختلف انتخاب ۹۶
شکل ۲۸- دقت روشهای مختلف انتخاب + NN معرفی شده روی داده‌های نویزی ۹۸
شکل ۲۹- مقایسه روشهای مختلف انتخاب خصیصه با طبقه بند پیشنهادی ۱۰۲
شکل ۳۰- مقایسه نتیجه‌ی GOR با استفاده از طبقه بند های مختلف ۱۰۳

فهرست اختصارات

ABBREVIATION	MEANING
AR	Auto Regressive
BMD	Bipolar Mode Disorder
DB	Davies Boildin
DB-FFR	Davies Bouldin Fast Feature Reduction
GOR	Greedy Overall Relevancy
MNN	Modified Nearest Neighbor
ONR	Overall Normalize Relevancy
R-N-R	Relevant-Non- Redundant

فصل اول

مقدمه

۱- مقدمه

۱-۱- انتخاب خصیصه

جمع آوری داده با انجام آزمایش کار بسیار ساده و راحتی است. داده ها با سرعت بی سابقه ای گرد آوری و ذخیره می شوند که بسیار بیشتر از سرعت پردازش انسان است بنابراین پیش پردازش این داده‌های تل انبار شده به منظور استخراج یک مجموعه داده کاراتر امر بسیار مهمی در مسائل یادگیری ماشین و داده کاوی به شمار می‌رود. انتخاب خصیصه از رایج ترین تکنیکهاست که کارایی آن در برخورد با مجموعه داده‌های بزرگ در بسیاری از مسائل یادگیری ماشین از قبیل: طبقه بندی، خوشه سازی و داده کاوی از سال ۱۹۷۰ ثابت شده است[1,2,3,4]. رشهای انتخاب خصیصه در دو دسته‌ی کلی یادگیری ماشین یعنی روشهای همره با ناظر و نیز روشهای بدون ناظر به کار برده شده‌اند اما در این رساله ما توجه خود را به مسائل همراه با ناظر (طبقه بندی) متمرکز کرده‌ایم که در آنها برچسب کلاس هر کدام از داده‌ها از ابتدا در مسئله مشخص است. پس یک مرور کلی بر روشهای انتخاب خصیصه در مسائل طبقه بندی کننده در این رساله انجام می‌شود. برای هر کدام از روشها مهمترین زمینه-های کاربردی بازنگری خواهد شد و در پایان یک روش جدید ارائه خواهد شد تا در حالیکه سعی دارد فواید روشهای قبلی را به ارث ببرد با دقت بهتری خصیصه‌های بهینه را گزینش کند.

انتخاب یک زیر مجموعه از خصیصه‌ها به دو موضوع برای بررسی.نیازمند است: ۱- یک الگوریتم و استراتژی جستجو کننده برای انتخاب خصیصه‌های کاندید و ۲- یک تابع هدف برای ارزیابی خصیصه‌های کاندید تا میزان خو بی این کاندیدها مشخص شود. از طرفی باید نشانه‌ای از مقدار ارزیابی شده توسط تابع هدف، به عنوان بازخورد به الگوریتم جستجو کننده باز گردانده شود تا در انتخاب کاندیداهای بعدی در نظر گرفته شود. الگوریتم‌های جستجوگربه طور کلی به سه دسته تقسیم می‌شوند: الف) الگوریتم‌های نمایی[۱]، ب) الگوریتم‌های ترتیبی[۲] و الگوریتم‌های تصادفی[۳]. و از طرفی الگوریتم‌های ترتیبی خود به دو گروه روش‌های پیشرونده[۴]

[1] exponential Algorithms
[2] Sequential Algorithms
[3] Randomized Algorithms
[4] Forward

و پسرونده[5] دسته بندی می‌شوند. روش‌های پسرونده کار خود را با مجموعه‌ای با حضورتمام خصیصه‌ها شروع می‌کند و سپس به طور ترتیبی خصیصه‌هایی که بیشترین کاهش در تابع هدف را موجب می‌شوند از این مجموعه حذف می‌کند. و روش‌های پیشرونده ابتدا مجموعه جواب را خالی فرض می‌کند و گام به گام در هر مرحله خصیصه‌ای را به جواب اضافه می‌کند که بیشترین تغییر را در مقدار تابع هدف موجب می‌شود.

بطور کلی توابع هدف مورد استفاده درروش‌های انتخاب خصیصه در دو دسته قرار می‌گیرند[32]: فیلترها[6] و رپرها[7]. در رویکرد فیلتر، انتخاب بهترین زیرمجموعه ازخصیصه‌هادر واقع یک مرحله پیش‌پردازش است و بعدا خصیصه‌ها برای طبقه بندی به الگوریتم‌های یادگیری ماشین داده می‌شوند. این روش بر اساس یک معیار از پیش تعیین شده می‌باشد که مستقل از کارایی خصیصه‌های انتخاب شده در افزایش دقت طبقه بند استفاده شده است. هر چند رویکرد فیلتر زمان اجرای کمتری لازم دارد اما این عیب را دارد که به طبقه بندی کننده مورد استفاده در فاز بعدی مسئله توجه نمی‌کند. در حالیکه در رویکرد رپر، همکنش بین مجموعه خصیصه انتخاب شده و طبقه بندی کننده به طور خاص مورد توجه روش جستجو در خصیصه‌ها قرار می‌گیرد. به عبارتی رپرها یک ماشین یادگیری خاص برای جستجو زیرمجموعه خصیصه مناسب بکار می‌برند. ماشین یادگیری روی یکی از زیرمجموعه خصیصه‌های موردنظر آموزش داده می‌شود و درجه دقت یادگیرنده در طبقه بندی داده‌ها به عنوان معیار ارزیابی زیرمجموعه ای از خصیصه‌های کاندید بازگردانده می‌شود. این مرحله برای همه‌ی زیرمجموعه‌های کاندید از خصیصه‌ها تکرار می‌شود تا جایی‌که بهترین زیرمجموعه خصیصه یافت شود. رویکرد رپر توسط Kohavi وJohn معرفی شد، و بطور کلی می‌تواند طبقه‌بندی کننده‌ای نهایی با دقت پیشگویی خوب تولید کند[32]؛ اما رویکرد رپر در مقایسه با روش فیلتر هزینه‌ی محاسباتی بیشتری دارد و احتمال داشتن مشکل انطباق بیش از حد[8] را افزایش می‌دهد[33,34]. بنابراین در بسیاری کاربردها، روش‌های فیلتر نسبت به رپر ترجیح داده‌می‌شوند.

معیار انتخاب خصیصه نقش حیاتی در روش‌های فیلتر بازی می‌کند. در مبحث طبقه‌بندی، این معیار سعی در اندازه‌گیری توانایی یک خصیصه یا یک مجموعه خصیصه برای جدا کردن کلاسهای مختلف دارد. تا به حال معیارهای مختلفی مانند معیارهای فاصله، وابستگی، و سازگاری برای انتخاب خصیصه بکار رفته‌است[35,36-37]. ولی این معیارها ممکن است به مقادیر قطعی داده‌های آموزشی بسیار حساس باشند؛ در نتیجه آنها به راحتی تحت تاثیر نویز[9] یا داده‌های بیرونی[10] قرار می‌گیرند. درحالیکه معیارهای اطلاعات، از قبیل انتروپی[11] یا

[5] Backward
[6] Filter approach
[7] wrapper
[8] Over-fitting
[9] Noise
[10] Outliers

اطلاعات متقابل، میزان اطلاعات یا عدم قطعیت یک خصیصه را برای طبقه‌بندی بررسی می‌کنند. بر تئوری اطلاعات Shannon[38, 39]، مدلی که مقدار اطلاعات را افزایش ندهد بی‌فایده تلقی می‌شود و انتظار نمی‌رود دقت پیشگویی آن بهتر از تنها یک حدس تصادفی باشد[40]. یکی از محاسن معیار میزان اطلاعات یک خصیصه این است که این معیار تنها به توزیع احتمال یک متغیر تصادفی وابسته‌است، نه به مقادیر قطعی آن. معیارهای اطلاعات بطور وسیع برای انتخاب خصیصه مورد استفاده قرار گرفته‌اند[42 ,41]، از قبیل بسیاری الگوریتم‌های یادگیری مانند $C_{4.5}$ [43]. یکی از این معیارها اطلاعات متقابل[12] می‌باشد.

۱-۲-ارزیابی یک زیر مجموعه از خصیصه‌ها

در همه کاربردها، به منظور ارزیابی یک زیر مجموعه از خصیصه‌ها میزان خوبی این زیر مجموعه توسط یک معیار معین به دست می‌آید. معیارهای ارزیابی به طور وسیع با توجه به وابستگی آنها به الگوریتم کاوشی که بعدا روی زیرمجموعه انتخاب شده عمل خواهد کرد در دو گروه دسته بندی می‌شوند: معیارهای مستقل[13] و معیارهای وابسته[14].

۱-۲-۱-معیارهای مستقل

معیارهای مستقل عموما در الگوریتم‌های فیلتر استفاده می‌شوند. این معیارها سعی دارند میزان خوبی یک خصیصه را با بهره‌گیری از ویژگیهای ذاتی داده‌های آموزشی و بدون در نظر گرفتن الگوریتم‌های کاوش اندازه گیری کنند.

برخی معیارهای مستقل عبارتند از مقیاس فاصله، مقدار اطلاعات، مقدار همبستگی و میزان سازگاری یا انطباق[5,6]. مقیاس فاصله با نامهای دیگری مثل مقدار جدایی، انحراف و کجی یا مقدار تشخیص پذیری نیز به کار برده شده است. از آنجا که در یک مسئله دو کلاسه هدف یافتن خصیصه‌ای است که کلاس‌ها را تا حد امکان از هم جدا کند، خصیصه X نسبت به خصیصه Y ارجح است اگربا در نظر گرفتن خصیصه X فاصله بیشتری بین احتمال شرطی کلاس‌ها ایجاد شود. یک خصیصه را غیر قابل تمیز دادن می‌گوییم اگر فاصله بین کلاس‌ها در حضور آن صفر شود.معیار مقدار اطلاعات معمولا میزان افزایش اطلاعات را معین می‌کنند. میزان افزایش تفاوت بین عدم قطعیت از پیش مشخص شده در داده‌ها وقتی خصیصه X را

[11] Entropy
[12] Mutual information
[13] Independent Criteria
[14] Dependency Criteria

استفاده می‌کنیم و عدم قطعیت پسین با حضور خصیصه X است. هرچه مقدار اطلاعات یک خصیصه بیشتر باشد مطلوب تر است پس در هر مرحله خصیصه با مقدار اطلاعات بیشتر انتخاب می‌شود.

معیارهایی نیز به عنوان اندازه‌ای از همبستگی و شباهت شناخته شده‌اند و در واقع توانایی پیش‌بینی کردن یک مقدار از مقادیر دیگر را اندازه می‌گیرند. در مسئله انتخاب خصیصه برای طبقه بندی ما به دنبال خصیصه‌هایی هستیم که بیشترین همبستگی را با برچسب کلاس هر داده دارد. و در مسئله انتخاب خصیصه برای خوشه‌سازی داده‌ها، میزان شباهت بین دو خصیصه مد نظر قرار می‌گیرد.

معیارهای سازگاری مشخصه‌های متفاوتی نسبت به معیارهای دکر شده دارند. این معیارها به داده‌های کلاس متکی هستند[6] و سعی می‌کنند مینیمم تعداد خصیصه‌ای را پیدا کنند که مطابق با وقتی که همه خصیصه‌ها را در نظر می‌گیریم بتواند کلاس‌ها را از هم جدا کند. عدم انطباق دو خصیصه به معنی این است که علیرغم داشتن مقدار یکسان در این خصیصه‌ها دو برچسب متفاوت بر اساس آنها انتخاب شود.

۱-۲-۲-معیارهای وابستگی

در رپرها از معیارهای همبستگی استفاده می‌شود و لازم است یک الگوریتم کاوش از قبل تعریف شده داشته باشند. رپرها اغلب کارایی بشتری دارندچون خصیصه‌هایی که برای این الگوریتم کاوش مناسب ترند را پیدا می‌کنند ولی در عوض هزینه محاسباتی زیادی دارند و در ضمن ممکن است برای الگوریتم‌های کاوش دیگر، کارایی آنها کاهش پیدا کند.

همه‌ی روشهای انتخاب خصیصه گفته شده به یک معیار یا شرط برای توقف کار نیاز دارند. معیار توقف معمولا یکی از موارد زیر است:

۱- جستجو کامل شده باشد،

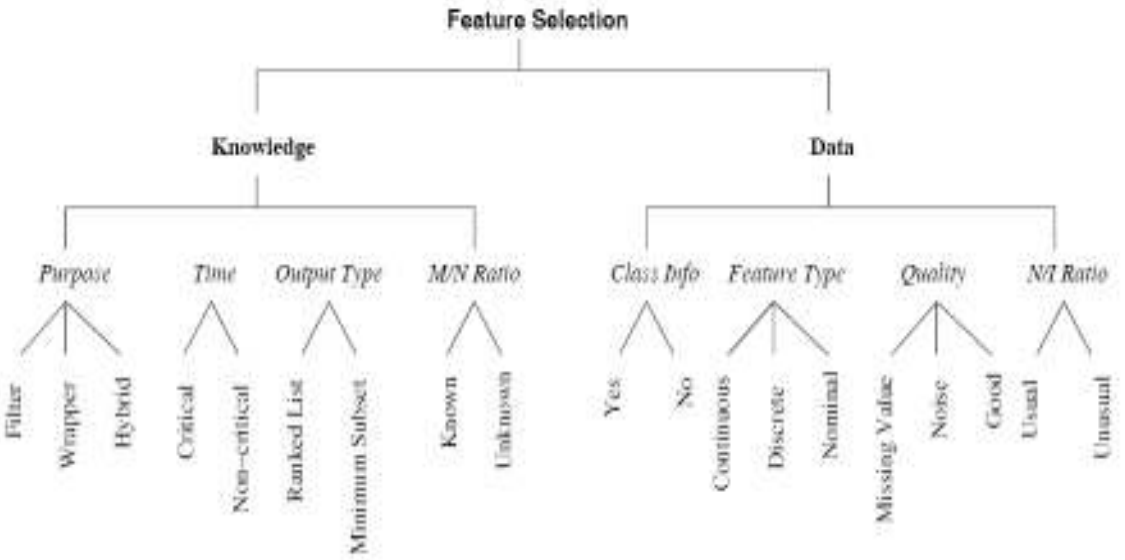
۲- یک حد بالایی مشخصی مرتفع شده است (به عنوان مثال به مینیمم تعداد خصیصه‌ها رسیده‌ایم و یا به ماکزیمم تعداد تکرار) ،

۳- دیگر اضافه یا حذف کردن خصیصه‌ی بعدی تاثیری بر کارایی مجموعه خصیصه انتخاب شده تاکنون ندارد و

۴- وقتی که یک زیر مجموعه‌ی به اندازه کافی خوب از خصیصه‌ها انتخاب شده باشد.

در نهایت پس از انتخاب زیر مجموعه ای از خصیصه‌ها با یکی از روشهای مورد نظر وقت آن است که این مجموعه را ارزیابی کنیم. یک راه ساده برای ارزیابی نتایج آن است که مستقیما

با استفاده از اطلاعاتی که از قبل راجع به داده‌ها داریم نتیجه را مقدار سنجی کنیم. اگر از قبل خصیصه‌های مرتبط و مطرح در جواب بهینه را بدانیم می‌توانیم این مجموعه معین را با مجموعه انتخاب شده مقایسه کنیم. اینکه اطلاعاتی درباره خصیصه‌هی غیر مرتبط و افزونه داشته باشیم هم می‌تواند مفید باشد. زیرا انتظار داریم این خصیصه‌ها انتخاب نشوند. البته در کاربردهای دنیای واقعی معمولا این دانش را از قبل راجع به خصیصه‌ها نداریم بنابراین باید کار را به طور غیر مستقیم دنبال کنیم، به این ترتیب که بر میزان تغییر در کارایی الگوریتم کاوش (طبقه بند یا خوشه ساز) وقتی خصیصه‌ها ی شرکت کننده را تغییر می‌دهیم نظارت کنیم. مثلا اگر نرخ خطای طبقه بندی را به عنوان مشخص کننده کارایی عمل کاوش انتخاب کنیم می‌توانیم به سادگی این نرخ را قبل و بعد از اضافه یا حذف یک خصیصه حساب کنیم.[5]. در نمودار ۱ یک قالب دسته بندی از روشهای انتخاب خصیصه آمده است.



شکل ۱ -دسته بندی روش‌های انتخاب خصیصه[1]

همان‌طور که می‌بینید دو فاکتور اصلی دانش از قبل مشخص شده و داده‌ها هستند. دانش خود شامل: هدف، زمان مورد نظر، نوع خروجی و نرخ $\frac{M}{N}$است که در آن M تعداد مورد انتظار برای زیر مجموعه ای که می‌خواهیم انتخاب کنیم و N تعداد کل مجموعه اصلی خصیصه-