



پایان نامه‌ی کارشناسی ارشد در رشته‌ی نرم‌افزار

روش‌های وزن‌دهی برای انتخاب ویژگی

به کوشش

یاسر تابنده

استاد راهنما

دکتر اشکان سامی

شهریور ماه 1390



به نام خدا

اظہارنامہ

اینجانب یاسر تابنده (870532) دانشجوی رشته ی مهندسی کامپیوتر گرایش نرم افزار دانشکده ی مهندسی اظہار می کنم کہ این پایان نامہ حاصل پژوهش خودم بوده و در جاهایی کہ از منابع دیگران استفادہ کرده ام، نشانی دقیق و مشخصات کامل آن را نوشته ام. همچنین اظہار می کنم کہ پژوهش و موضوع پایان نامہ ام تکراری نیست و تعہد می نمایم کہ بدون مجوز دانشگاه دستاوردهای آن را منتشر ننمودہ و یا در اختیار غیر قرار ندهم. کلیہ حقوق این اثر مطابق با آیین نامہ مالکیت فکری و معنوی متعلق بہ دانشگاه شیراز است.

نام و نام خانوادگی:

تاریخ و امضا:

به نام خدا

روش های وزن دهی برای انتخاب ویژگی

به کوشش

یاسر تابنده

پایان نامه

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

در رشته ی:

مهندسی کامپیوتر گرایش نرم افزار

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی شده توسط کمیته پایان نامه با درجه:

..... دکترا اشکان سامی، استادیار بخش مهندسی کامپیوتر (رئیس کمیته)

..... دکترا هادی صدرالدینی، دانشیار بخش مهندسی کامپیوتر

..... دکترا رضا ثامنی، استادیار بخش مهندسی کامپیوتر

شهریور ماه 90

تقدیم به پدر، مادر و همسر مهربانم

با تشکر و سپاس فراوان از کلیه عزیزانی که مرا در انجام این پژوهش یاری کردند.

چکیده

روش های وزن دهی برای انتخاب ویژگی

توسط

یاسر تابنده

انتخاب ویژگی یکی از مهم ترین اقدامات پیش پردازش در عملیات داده کاوی می باشد. با انجام این مرحله از فرایند، حجم داده های پردازشی کمتر می شود، عملیات داده کاوی سریع تر و دقت الگوریتم های یادگیری بیشتر می شود. روش های انتخاب خصیصه از لحاظ نحوه انتخاب به دو نوع انتخاب مجموعه ای و رتبه بندی خصیصه ها طبقه بندی می شوند. در این پژوهش مساله انتخاب ویژگی و مهم ترین روش های ارائه شده که از طریق رتبه بندی خصیصه ها به انتخاب خصیصه می پردازند مورد بررسی قرار می گیرد و همچنین چند روش جدید برای وزن دهی و رتبه بندی ویژگی ها ارائه می شود. مهم ترین کارهای ارائه شده در این پژوهش عبارتند از:

- روشی جدید برای محاسبه فاصله بین دو ویژگی غیر عددی
- روشی سریع برای محاسبه سریع تر فاصله بین نمونه های داده
- بهبود الگوریتم ReliefF برای داده های چند کلاسه
- روشی ترکیبی برای بهبود الگوریتم های ReliefF و Gain Ratio
- بهبود روش Chi-Square برای ویژگی های غیر عددی با تعداد مقادیر زیاد
- ارائه الگوریتم وزن دهی جدید مبتنی بر فاصله برای ویژگی های عددی
- ارائه یک الگوریتم wrapper تصادفی برای انجام رتبه بندی شامل دو روش وزن دهی
- ارائه روشی wrapper برای انجام همزمان انتخاب مجموعه ای و وزن دهی ویژگی بر اساس الگوریتم زنبورها

روش های ارائه شده در این پژوهش بر روی داده های استاندارد UCI آزمایش و با سایر الگوریتم ها و روش های مطرح مقایسه شده اند، همچنین برخی روش ها در مسابقات داده کاوی به عنوان پیش پردازش استفاده شده اند که نتیجه آن کسب رتبه های بالا در این مسابقات می باشد.

فهرست مطالب

عنوان	صفحه
فصل 1- مقدمه.....	8
1-1- داده کاوی (Data mining).....	8
2-1- انتخاب ویژگی (Feature Selection).....	10
1-2-1- تعریف ویژگی (Feature).....	10
2-2-1- انتخاب ویژگی چیست؟.....	11
3-2-1- اهداف انتخاب ویژگی.....	12
4-2-1- انواع ویژگی ها.....	12
3-1- طبقه بندی الگوریتم های انتخاب ویژگی.....	13
1-3-1- شمای اصلی الگوریتم های انتخاب ویژگی.....	13
2-3-1- مقایسه روش های wrapper و filter.....	15
1-2-3-1- روش های wrapper.....	15
2-2-3-1- روش های filter.....	17
3-3-1- طبقه بندی کلی.....	17
4-1- در مورد این پژوهش.....	18
فصل 2- مروری بر تحقیقات پیشین.....	22
1-2- روش های filter.....	22
1-1-2- روشهای مبتنی بر فاصله.....	22
1-1-1-2- Relief (Kira and Rundell 1992).....	22
2-1-1-2- ReliefF (Kononenko, 1994).....	24
3-1-1-2- I-Relief (Sun, 2006).....	24
4-1-1-2- Constraint Score(Zhang et al., 2007).....	24
2-1-2- معیار های بر اساس هم بستگی (Correlation).....	25
1-2-1-2- Information Gain (Quinlan, 1986).....	25

25	Gain Ratio (Quinlan, 1993) -2-2-1-2
26	mRMR (Ding and Peng 2003) -3-2-1-2
26	الگوریتم های آماری -3-1-2
27	T Test -1-3-1-2
27	Fisher Score (Duda and Stork, 2001) -2-3-1-2
28	روش های wrapper -2-2
28	SVM-RFE (Guyon et al., 2002) -1-2-2
29	RMS-Liknon (Lai et al., 2006) -2-2-2
29	GA-KNN (Pei et al. , 1999) -3-2-2
31	فصل 3- کارهای ارائه شده در این پژوهش
31	1-3- روش انجام کار
31	1-1-3- ابزارهای استفاده شده
31	2-1-3- Data Set های استفاده شده برای تست روش ها
31	1-2-1-3- داده های استاندارد UCI
32	3-1-3- نحوه و معیار انجام تست
33	2-3- کارهای ارائه شده در این پژوهش
34	1-2-3- روش هایی برای انجام بهتر روش های مبتنی بر فاصله
35	1-1-2-3- روشی بهتر برای محاسبه فاصله دو ویژگی غیر عددی در داده های binary
35	1-1-1-2-3- روش موجود
35	2-1-1-2-3- روش ارائه شده
36	3-1-1-2-3- نتایج
37	2-1-2-3- روشی برای محاسبه سریع فاصله بین دو instance در داده های عددی
37	1-2-1-2-3- روش ارائه شده
38	2-2-1-2-3- ارزیابی روش ارائه شده روی داده های UCI
39	2-2-3- روش های Filter
40	1-2-2-3- بهبود الگوریتم ReliefF برای داده های Multiclass
40	1-1-2-2-3- تعریف Margin جدید
40	2-1-2-2-3- مقایسه دو روش
42	2-2-2-3- ارائه یک روش ترکیبی برای الگوریتم های Relief و GR
42	1-2-2-2-3- ارائه روش ترکیبی جدید
42	2-2-2-2-3- نتیجه اجرای الگوریتم بر روی داده های UCI

44 نتیجه اجرای الگوریتم در مسابقات داده کاوی
46 بهبود الگوریتم Chi-Square برای ویژگی های غیر عددی با تعداد مقادیر زیاد
46 بهبود الگوریتم
47 نتیجه اجرای الگوریتم بر روی داده های UCI
47 نتیجه اجرای الگوریتم در مسابقات داده کاوی
49 ارائه یک الگوریتم جدید مبتنی بر فاصله
49 معایب روش های موجود
49 ارائه روش جدید
50 نتایج ارزیابی الگوریتم ارائه شده
51 روش های Wrapper
51 ارائه روش wrapper تصادفی برای رتبه بندی خصیصه ها
52 وزن دهی از طریق میانگین گیری
53 وزن دهی از طریق Regression
54 مقایسه دو روش در داده های UCI
55 مقایسه دو روش با ranker های دیگر
56 انجام همزمان انتخاب مجموعه ای و رتبه بندی بر اساس الگوریتم زنبورها با روش wrapper
56 نحوه وزن دهی ویژگی ها
57 نتایج feature subset selection
57 نتایج feature ranking
49	فصل 4- خلاصه و نتیجه گیری
49 خلاصه
49 نتیجه گیری
50 پیشنهادات برای پژوهشهای آینده
60 منابع و ماخذ

فهرست جدول‌ها

عنوان	صفحه
جدول 1 مشخصات برخی از داده های باینری UCI استفاده شده در این پژوهش	32
جدول 2 مشخصات برخی از داده های چند کلاسه UCI استفاده شده در این پژوهش	32
جدول 3 دقت اجرای الگوریتم KNN توسط روش موجود و روش جدید	36
جدول 4 نتیجه معیار AATRF بر روی داده های باینری توسط الگوریتم ReliefF و کلاسیفیر KNN	36
جدول 5 ارزیابی الگوریتم ReliefF در حالت معمولی و در حالت سریع	38
جدول 6 مقایسه الگوریتم ReliefF در حالت معمولی با ReliefF در حالت Margin ارائه شده	41
جدول 7 مقایسه الگوریتم ترکیبی با الگوریتم های ReliefF و GR توسط KNN بر روی داده های باینری	42
جدول 8 مقایسه الگوریتم ترکیبی با الگوریتم های ReliefF و GR توسط KNN بر روی داده های چند کلاسه	43
جدول 9 مقایسه الگوریتم ترکیبی با الگوریتم های ReliefF و GR توسط Naive Bayes بر روی داده های باینری	43
جدول 10 مقایسه الگوریتم ترکیبی با الگوریتم های ReliefF و GR توسط Naive Bayes بر روی داده های چند کلاسه	43
جدول 11 نتایج AATRF بر روی داده های UCI مقایسه Chi-Square معمولی با Chi-Square بهبود داده شده	47
جدول 12 نتایج اجرای KNN و Naive Bayes بر روی داده های عددی UCI و مقایسه الگوریتم ارائه شده RegD	50
جدول 13 نتیجه اجرای الگوریتم از دو روش در تکرارهای متفاوت	54
جدول 14 مقایسه دو روش SWR با ranker های مطرح	55
جدول 15 مقایسه BSS با full feature set	57
جدول 16 مقایسه BFR با ReliefF و Gain Ratio	57

فهرست تصویرها

صفحه	عنوان
11	شکل 1 مثال جدول اطلاعاتی برای ویژگی ها
11	شکل 2 مثال انتخاب ویژگی از بین ویژگی های مربوط به افراد
14	شکل 3 شمای کلی یک الگوریتم انتخاب خصیصه
15	شکل 4 شمای کلی یک روش wrapper
17	شکل 5 شمای کلی یک روش filter
18	شکل 6 طبقه بندی کلی الگوریتم های انتخاب ویژگی
23	شکل 7 نمونه ای از وزن دهی Relief در فضای دو بعدی
23	شکل 8 الگوریتم Relief
27	شکل 9 وزن دهی در t-score
28	شکل 10 الگوریتم SVM-RFE
29	شکل 11 الگوریتم RMS-Liknon
29	شکل 12 الگوریتم GA-KNN
37	شکل 13 نتیجه مسابقه KDD Cup 2010
44	شکل 14 نمودار مقایسه ای الگوریتم GR-Relief با ReliefF و Gain Ratio
45	شکل 15 نتیجه مسابقه UCSD 2010
45	شکل 16 نتیجه مسابقه SDM 2010
48	شکل 17 نتیجه مسابقه PAKDD 2010
48	شکل 18 نتیجه مسابقه Predict Grant Application
53	شکل 20 الگوریتم SWR-Regression

فصل اول

فصل 1 - مقدمه

در این قسمت انتخاب ویژگی به عنوان یکی از مهم ترین مراحل در انجام عملیات داده کاوی مورد بحث قرار می گیرد و روش وزن دهی که یکی از روش های معمول و کارآمد در این زمینه است به صورت جزئی شرح داده می شود.

1-1 - داده کاوی (Data mining)

داده هر موجودیتی است که بتواند توسط رایانه مورد پردازش قرار گیرد و امروزه به صورت ها و فرمت های گوناگون و با ابعاد و حجم های متفاوت توسط افراد به کار برده می شوند. حقایق، قوانین و ارتباطات بین این داده ها می تواند به اطلاعات منجر شود و خود این اطلاعات می تواند تبدیل به دانش شود. به طور مثال، داده ها و اطلاعات مربوط به فروش یک شرکت و مشتری های آن، می تواند به صاحبان آن شرکت دانشی در مورد رفتار مشتریان، میزان فروش و پیش بینی فروش آینده شرکت بدهد.

مجموعه فرایند هایی که هدف آنها جمع آوری داده ها و خلاصه کردن آنها به صورت اطلاعاتی که بتواند به گونه ای مفید مورد استفاده قرار گیرد، باعث بوجود آمدن شاخه ای جدید در علوم کامپیوتر به نام داده کاوی شده است.

داده کاوی با علوم مختلفی مانند مدیریت پایگاه داده، آمار و هوش مصنوعی رابطه دارد و از آنها در فرایند های خود استفاده می کند. اهداف زیادی در داده کاوی می تواند دنبال شود از جمله این اهداف می توان به موارد زیر اشاره کرد:

- پیش بینی: بسته به داده ها، اطلاعات و آمار موجود می توان از داده کاوی برای پیش بینی آمار آینده استفاده نمود.
- کلاس بندی: با استفاده از داده کاوی می توان داده ها و نمونه های مختلف را طی یک سری فرایندها به شکل های گوناگون طبقه بندی کرد.
- خوشه بندی: یک هدف دیگر داده کاوی خوشه بندی داده ها بر حسب خصوصیات آنهاست.

- کاوش قوانین وابستگی: پیدا کردن برخی قوانین مخفی بین اطلاعات و استفاده از آنها
دیگر هدف داده کاوی است

داده کاوی کاربردهای زیادی در بسیاری از زمینه ها دارا می باشد که از جمله این کاربردها می توان برخی موارد زیر را مثال زد:

- علم زیست شناسی و پردازش ژن ها برای شناخت بهتر خصوصیات جانداران
- علم شیمی و کشف خصوصیات مواد نا شناخته بر حسب ویژگی های ظاهری
- علوم اقتصادی و مالی برای ارزیابی مشتریان ، سیستم های تصمیم گیری و پیش بینی های مالی
- پردازش زبانهای طبیعی، ترجمه زبان ها، تشخیص گفتار و تشخیص نوشتار، خلاصه کردن متون
- پردازش تصاویر، پردازش سیگنالهای الکتریکی

و ده ها زمینه دیگر که داده کاوی را به یک علم بسیار کاربردی و مفید تبدیل کرده اند. به طور کلی داده کاوی شامل ریز فرایندهای زیادی می باشد که در نهایت منجر به کشف دانش می شود. عموماً این ریز فرایندها شامل موارد زیر می شود:

1. جمع آوری داده ها: در این مرحله سعی می شود بسته به هدف داده کاوی، داده های مورد نیاز به صورت فرمت های مورد نیاز و یکجا برای فرایند های بعدی جمع آوری شود.

2. پیش پردازش: در این مرحله داده های جمع آوری شده به فرمت های مورد نیاز ابزار داده کاوی تبدیل می شوند از جمله:

- دور ریختن داده های ناخواسته
- جایگزین کردن ویژگی های بدون مقدار
- انتخاب ویژگی
- تبدیل داده ها

3. پردازش اصلی: بر حسب هدف فرایند، این مرحله شامل یکی از موارد زیر می شود که به آنها الگوریتم های یادگیری هم گفته می شود:

- رگرسیون (Regression): بیشتر برای پیش بینی های عددی به کار می رود.

- کلاس بندی (Classification): بیشترین کاربرد داده کاوی که برای کلاس بندی داده ها به کار می رود و یک یا چند برچسب به هر نمونه از داده ها داده می شود.
- خوشه بندی (Clustering): همانند کلاس بندی با این تفاوت که داده ها برچسب ندارند و بر حسب ویژگیهایشان طبقه بندی می شوند.
- کشف قواعد وابستگی (Assosiation Rule Mining): برای مشخص کردن برخی قواعد مخفی در داده ها بکار می رود.

4. ارزیابی: در این مرحله نتایج بدست آمده بر حسب معیارهایی مشخص مورد ارزیابی قرار می گیرد.

موفقیت هر عملیات داده کاوی به نحوه صحیح انجام زیر فرایند های آن بستگی دارد، مثلاً برای عملیات کلاس بندی تنها انتخاب یک الگوریتم خوب کلاس بندی کافی نیست، بلکه انجام صحیح عملیات پیش پردازش می تواند نتیجه این الگوریتم را به میزان خیلی زیاد بهبود دهد.

1-2- انتخاب ویژگی (Feature Selection)

همانگونه که در بخش قبل دیدیم، انتخاب ویژگی یکی از مهم ترین و گاهی اوقات مهم ترین مرحله در گام پیش پردازش مجموعه فرایندهای داده کاوی است. در این قسمت به تعریف جامع تری از انتخاب ویژگی می پردازیم.

1-2-1- تعریف ویژگی (Feature)

به خصوصیات یک نمونه از داده ها که در مورد نمونه های مختلف تغییر می کند ویژگی گفته می شود.

به طور خاص می توان گفت ستون های یک جدول اطلاعاتی همان ویژگی های آن و ردیف های آن نمونه ها (Instance) هستند.

TRS_DT	TRS_TYP_CD	REF_DT	REF_NUM	CO_CD	GDS_CD	QTY	UT_CD	UT_PRIC
21/05/93	00001	04/03/93	25119	10002J	001M	10	CTN	22.000
21/05/93	00001	05/05/93	25124	10002J	032J	200	DOZ	1.370
21/05/93	00001	05/05/93	25124	10002J	033Q	500	DOZ	1.000
21/05/93	00001	13/05/93	25217	10002J	024K	5	CTN	21.000
21/05/93	00001	13/05/93	25216	10026H	006C	20	CTN	69.000
21/05/93	00001	13/05/93	25216	10026H	008Q	10	CTN	114.000
21/05/93	00001	14/05/93	25232	10026H	006C	10	CTN	69.000
21/05/93	00001	14/05/93	25235	10027E	003A	5	CTN	24.000
21/05/93	00001	14/05/93	25235	10027E	001M	5	CTN	24.000
21/05/93	00001	22/04/93	24974	10035E	009F	50	CTN	118.000
21/05/93	00001	27/04/93	25033	10035E	015A	375	GRS	72.000
21/05/93	00001	20/05/93	25313	10041Q	010F	10	CTN	26.000
21/05/93	00001	12/05/93	25197	10054R	002E	25	CTN	24.000

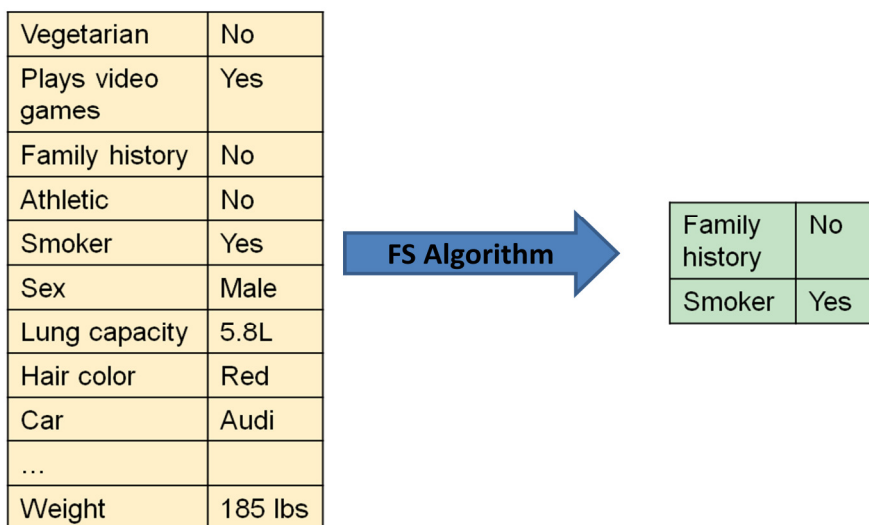
شکل 1 مثال جدول اطلاعاتی برای ویژگی ها

1-2-2- انتخاب ویژگی چیست؟

انتخاب ویژگی که گاهی اوقات از آن به نام انتخاب متغیر و کم کردن ویژگی هم یاد می شود عبارتست از سری عملیاتی که منجر به انتخاب چند ویژگی از مجموعه ویژگی های مربوط به داده ها می شود به صورتی که اطلاعات زیادی از این داده ها از بین نرود، یعنی یک تبدیل از داده ها با ابعاد زیاد به داده های با ابعاد کمتر.

فرض کنید یک سری اطلاعات در مورد افراد و ویژگی های آنها داشته باشیم که این ویژگی ها شامل قد، وزن، فشار خون، رنگ مو، سن، جنسیت، رنگ چشم، میزان استعمال دخانیات، ورزش، شغل و ... باشد. اگر هدف دسته بندی این افراد بر حسب میزان احتمال دچار شدن این افراد به بیماری ریه باشد، به طور یقین برخی ویژگیها مثل ورزش، میزان استعمال دخانیات و فشار خون بیشتر مفید هستند تا ویژگیهایی مثل رنگ مو یا شغل.

یا در مورد مثال زیر



شکل 2 مثال انتخاب ویژگی از بین ویژگی های مربوط به افراد

یک الگوریتم انتخاب ویژگی ممکن است از بین ویژگی های لیست شده فقط دو ویژگی بالا را انتخاب کند.

1-2-3- اهداف انتخاب ویژگی

انتخاب ویژگی در بسیاری موارد باعث بهبود فرایند داده کاوی می شود. از جمله اهدافی که از انتخاب ویژگی دنبال می شود می توان به موارد زیر اشاره کرد:

- کم کردن حجم داده ها: با انجام انتخاب ویژگی حجم زیادی از داده ها از فرایند داده کاوی کنار گذاشته می شود و این خود باعث تسهیل در فرایندهای بعدی می شود.
- تسریع عملیات: عموماً اکثر الگوریتم های داده کاوی که عملیات یادگیری را انجام می دهند به طور مستقیم و گاهی اوقات وابستگی نمایی با تعداد ویژگی های نمونه داده ها دارند و هر چه تعداد این ویژگی ها کمتر باشد سرعت پردازش آنها بیشتر می شود. پس انجام یک ریز فرایند انتخاب ویژگی قبل از این الگوریتم ها به طور حتم سرعت آنها را افزایش می دهد.
- بهبود دقت الگوریتمهای داده کاوی: وجود ویژگی های بی ربط و اضافی، به طور ناخواسته باعث ایجاد برخی اشتباهات و نقایص در الگوریتم های یادگیری که می خواهند از این ویژگی ها استفاده کنند می شود که این خود دقت این الگوریتم ها و کارایی آنها را به میزان قابل توجهی کاهش می دهد.
- فهم بهتر نتایج: خروجی بسیاری الگوریتم های یادگیری بر حسب رابطه ویژگی هاست، به طور مثال الگوریتم های درخت تصمیم گیری (Decision Tree) بر اساس ویژگی ها درخت تصمیم گیری را تشکیل می دهند، پس وجود ویژگی هایی که ربط بیشتری به موضوع و هدف الگوریتم دارند فهم این نتایج را بهتر می کند.

1-2-4- انواع ویژگی ها

به طور کلی ویژگی های موجود در داده ها به سه دسته تقسیم می شوند:

1. ویژگی های بی ربط: ویژگی هایی هستند که ربطی به هدف داده کاوی ندارند و هیچ اطلاعات خاصی به دست نمی دهند و در مواردی باعث کم شدن کارایی الگوریتم ها می شوند.

2. ویژگی های اضافه: این ویژگی ها به طور مستقیم بی ربط نیستند ولی ارتباط زیادی با سایر ویژگی ها دارند، مثل ویژگی هایی که وابستگی تابعی با یک یا چند ویژگی دیگر دارند.
3. ویژگی های مربوط: این ویژگی ها هدف اصلی الگوریتم های انتخاب ویژگی هستند و ارتباط زیادی با هدف الگوریتم های داده کاوی دارند.

1-3- طبقه بندی الگوریتم های انتخاب ویژگی

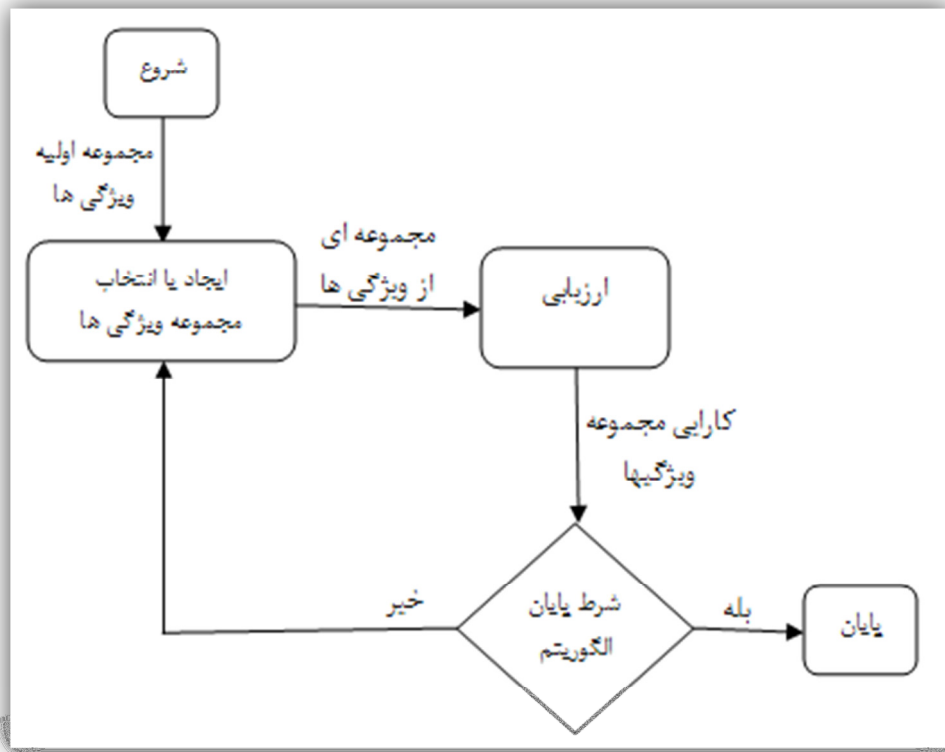
بسته به وجود یا عدم وجود برچسب بر روی نمونه های داده می توان الگوریتم های انتخاب ویژگی را به سه دسته تقسیم می شوند:

1. الگوریتم های Supervised: الگوریتم هایی هستند که از طریق رابطه بین ویژگی ها و برچسب آنها (Class Label) به انتخاب ویژگی می پردازند. این الگوریتمها اکثرا برای اهداف کلاس بندی و پیش بینی استفاده می شوند.
2. الگوریتم های Semi-Supervised: در این الگوریتم ها تعداد محدودی از نمونه ها دارای برچسب و حجم زیادی از آنها بدون برچسب می باشد. کاربرد این الگوریتم ها هم مثل دسته قبلی عموما برای کلاس بندی و پیش بینی از طریق Regression می باشد.
3. الگوریتم های Unsupervised: این الگوریتم ها در مورد داده هایی به کار می رود که همگی فاقد برچسب هستند. این الگوریتم ها برای اهداف خاص مثل خوشه بندی به کار می روند.

هدف اصلی در این پژوهش الگوریتم های supervised هستند که برای عمل کلاس بندی به کار می روند. این الگوریتم ها حائز اهمیت بیشتری نسبت به سایر الگوریتم ها که اکثرا کاربرد موردی دارند، می باشد.

1-3-1- شمای اصلی الگوریتم های انتخاب ویژگی

یک الگوریتم Feature Selection به طور کلی شامل مراحل زیر می باشد:



شکل 3 شمایی کلی یک الگوریتم انتخاب خصیصه

1. ایجاد یا انتخاب مجموعه ویژگی‌ها: در این مرحله، زیر مجموعه ای از ویژگی‌ها برای ارزیابی ایجاد می‌شود. شروع می‌تواند با کل ویژگی‌ها، بدون هیچ ویژگی یا یک زیرمجموعه احتمالی باشد و بعد از آن ویژگی‌ها می‌توانند از زیرمجموعه حذف، اضافه یا هم حذف و هم اضافه شوند.
2. مرحله ارزیابی: در این مرحله ویژگی‌های انتخاب شده در مرحله قبل ارزیابی می‌شوند، بر اساس نحوه ارزیابی، می‌توان روش‌های انتخاب ویژگی را به سه دسته wrapper و filter و embedded تقسیم کرد. در روش‌های wrapper ارزیابی توسط خود classifier که می‌خواهد این ویژگی‌ها را مورد استفاده قرار دهد صورت می‌گیرد ولی در روش‌های filter این ارزیابی توسط معیارهایی خاص صورت می‌گیرد. در روش‌های embedded، انتخاب ویژگی بخشی از الگوریتم یادگیری است. این روش‌ها در بخش بعد شرح داده خواهد شد.
3. چک کردن شرط پایان الگوریتم: در این مرحله بر حسب میزان تکرار الگوریتم یا رسیدن به بهترین زیرمجموعه از ویژگی‌ها، پایان الگوریتم چک می‌شود، در صورت