

**دانشگاه آزاد اسلامی**

**واحد قزوین**

دانشکده برق،کامپیوتر و فناوری اطلاعات، گروه فناوری اطلاعات

پایان‎نامه برای دریافت درجه کارشناسی ارشد “M.Sc”

رشته مهندسی فناوری اطلاعات، گرایش تجارت الکترونیک

**عنوان:**

یک روش­ تکاملی چند هدفه به منظور کشف انجمن در شبکه­های اجتماعی

**استاد راهنما :**

دکتر ...

**استاد مشاور:**

دکتر ...

**نگارش:**

.... شفاف

زمستان 1934





معاونت پژوهش و فناوری

بنام خدا

**منشور اخلاق پژوهش**

با یاری از خداوند سبحان و اعتقاد به این‌که عالم محضر خداست و همواره ناظر بر اعمال انسان و به منظور پاس داشت مقام بلند دانش و پژوهش و نظر به اهمیت جایگاه دانشگاه در اعتلای فرهنگ و تمدن بشری، ما دانشجویان و اعضاء هیأت علمی واحدهای دانشگاه آزاد اسلامی متعهد می‌گردیم اصول زیر را در انجام فعالیت‌های پژوهشی مد نظر قرار داده و ار آن تخطی نکنیم:

1- **اصل برائت**: التزام به برائت‌جویی از هر گونه رفتار غیر حرفه‌ای و اعلام موضع نسبت به کسانی که حوزه علم و پژوهش را به شائبه‌های غیر علمی می‌آلایند.

2- **اصل رعایت انصاف و عدالت**: تعهد به اجتناب از هر گونه جانب‌داری غیر علمی و حفاظت از اموال، تجهیزات و منابع در اختیار.

3- **اصل ترویج**: تعهد به رواج دانش و اشاعه نتایج تحقیقات و انتقال آن به همکاران علمی و دانشجویان به غیر از مواردی که منع قانونی دارد.

4- **اصل احترام**: تعهد به رعایت حریم‌ها و حرمت‌ها در انجام تحقیقات و رعایت جانب نقد و خوداری از هر گونه حرمت‌ شکنی.

5- **اصل رعایت حقوق**: التزام به رعایت حقوق پژوهشگران و پژوهیدگان (انسان، حیوان و نبات) و سایر صاحبان حق.

6- **اصل رازداری**: تعهد به صیانت از اسرار و اطلاعات محرمانه افراد، سازمان‌ها و کشور و کلیه افراد و نهادهای مرتبط با تحقیق.

7- **اصل حقیقت‌جویی**: تلاش در راستای پی‌جویی حقیقت و وفاداری به آن و دوری از هر گونه پنهان‌سازی حقیقت.

8- **اصل مالکیت مادی و معنوی**: تعهد به رعایت کامل حقوق مادی و معنوی دانشگاه و کلیه همکاران پژوهش.

**9- اصل منافع ملی**: تعهد به رعایت مصالح ملی و در نظر داشتن پیشبرد و توسعه کشور در کلیه مراحل پژوهش.

 معاونت پژوهشی فرم شماره 26

دانشگاه آزاد اسلامی قزوین

**تعهدنامه اصالت رساله یا پایان نامه**

اینجانب .............................. دانش­آموخته مقطع کارشناسی ارشد ناپیوسته در رشته ....................................... که در تاریخ ............. از پایان خود تحت عنوان ............................................................................................

........................................................................................................................................................................

با کسب نمره ....... ودرجه ........... دفاع نموده ام بدینوسیله متعهدمی شوم:

1) این پایان نامه / رساله حاصل تحقیق وپژوهش انجام شده توسط اینجانب بوده ودر مواردي که از دستاوردهاي علمی وپژوهشی دیگران (اعم از پایان نامه، کتاب، مقاله و ... ) استفاده نموده ام، مطابق ضوابط ورویه موجود، نام منبع مورد استفاده وسایر مشخصات آن را درفهرست مربوطه ذکر و درج کرده ام.

2) این پایان نامه / رساله قبلا" براي دریافت هیچ مدرك تحصیلی ( هم سطح، پائین تریا بالاتر) درسایر دانشگاه ها ومؤسسات آموزش عالی ارائه نشده است.

3) چنانچه بعد از فراغت از تحصیل، قصد استفده و هرگونه بهره برداري اعم از چاپ کتاب، ثبت اختراع و ... از این پایان نامه داشته باشم، از حوزه معاونت پژوهشی واحد مجوزهاي مربوطه را اخذ نمایم.

4) چنانچه در هر مقطع زمانی خلاف موارد فوق ثابت شود، عواقب ناشی ازآن را می پذیرم و واحد دانشگاهی مجاز است با اینجانب مطابق ضوابط ومقررات رفتار نموده و درصورت ابطال مدرك تحصیلی ام هیچ گونه ادعایی نخواهم داشت.

نام و نام خانوادگی تاریخ: امضاء

فرم شماره 18 (صورتجلسه دفاع)

تقدیم به

تقدیر و تشکر

چکیده فارسی

فهرست مطالب

[1-فصل اول: کلیات 1](#_Toc434767505)

[1-1 مقدمه 2](#_Toc434767506)

[1-2 نظریه گراف در شبکه‎های اجتماعی 3](#_Toc434767507)

[1-3 بیان مسئله 5](#_Toc434767508)

[1-4 اهداف و نوآوری‎های پژوهش 6](#_Toc434767509)

[2- فصل دوم: پیشینه‎ی تحقیق 9](#_Toc434767510)

[2-1 مبانی نظری پژوهش 10](#_Toc434767511)

[2-1-1 داده‎کاوی 10](#_Toc434767512)

[2-1-2 انتخاب ویژگی 11](#_Toc434767513)

[2-1-3 طبقه‌بندی روشهای مختلف انتخاب ویژگی 14](#_Toc434767514)

[2-1-4 یادگیری ماشین و طبقهبندی 18](#_Toc434767515)

[2-1-5 الگوریتم‎های فرامکاشفه‎ای زیستی 21](#_Toc434767516)

[2-2 مروری بر پژوهش‎های پیشین 22](#_Toc434767517)

فهرست شکل‎ها

[شکل (1- 1): نمونهای از گراف جهت دار یک شبکه 4](#_Toc434767947)

[شکل (1- 2): نمونهای از گراف بدون جهت یک شبکه 5](#_Toc434767948)

[شکل (2- 1): دسته بندی کلی انواع طبقه‎بندها 20](#_Toc434767958)

فهرست جدول‎ها

[جدول (1- 1) : انواع توابع ارزیابی و ویژگی‌های آنها 20](#_Toc434768363)

# فصل اول: کلیات

## 1-1 مقدمه

در دهه اخیر مفهوم شبکه­های اجتماعی برخط[[1]](#footnote-1) نظیر فیسبوک و گوگل­پلاس به یکی از مهمترین خدمات­دهندگان سامانه‎های ارتباطی به منظور انتشار و به اشتراک‎گذاری اطلاعات کاربران خود تبدیل شده است. از این طریق تحلیل شبکه­های اجتماعی به یکی از مهمترین ابزارها در بسیاری از کاربردهای هوشمندسازی تجارت الکترونیک نظیر شناسایی مشتریان مستعد و بازار بالقوه و حتی شاخص­های باارزش‎تر و پراهمیت بر اساس جذابیتشان تبدیل شده است [[1](#_ENREF_1)] [[2](#_ENREF_2)].

شبکه­های اجتماعی موضوعات پویای گسترده­ای را در برمی­گیرد که در طول زمان به سرعت در حال تغییر و رشد است. افزودن ارتباطات جدید در هر شبکه اجتماعی نشان دهنده‎ی وجود برهمکنش­های جدید در ساختار شبکه اجتماعی است [[3](#_ENREF_3)].

بسیاری از پایگاه­های اینترنتی ارائه‎دهنده‎ی شبکه­های اجتماعی برای هر کاربر امکان اضافه کردن کاربران دیگر را به فهسرت دوستان خود برای ایجاد یک شبکه گفتگو فراهم می­کنند. تحقیقات نشان داده است که کاربران علاوه بر ایجاد ارتباط با دوستان برخط، به آن دسته از دوستانی که می­دانند برخط نیستند نیز پیام ارسال می­کنند. افزون بر این درصد بالایی از افراد درشبکه‎های اجتماعی به دنبال برقراری ارتباط با دوستان جدید در شبکه­های اجتماعی نیز می­گردند [[4](#_ENREF_4)].

به عنوان مثال بسیاری از کاربران در شبکه­های پرطرفداری مانند فیسبوک[[2]](#footnote-2) و مای­اسپیس[[3]](#footnote-3) بیشتر با کاربرانی که می­دانند برخط نیستند ارتباط برقرار می­کنند. از سوی دیگر تحقیقات بر روی مؤسسات دست اندرکار شبکه­های اجتماعی نشان می­دهد که کاربران در یک زمینه مشترک علاقه مند به یافتن ارتباطات ارزشمندتری هستند که پیشتر با آن آشنایی نداشتند، افزون بر آن برای درارتباط ماندن با دوستان نزدیک به اموری که کمتر مورد علاقه آن­هاست متصل می­شوند. با توجه به اندازه بزرگ پایگاه­های شبکه­های اجتماعی، پیدا کردن ارتباطات شناخته شده و دوستان جدید مورد علاقه برای معرفی به یک کاربر دو چالش اساسی هستند [[4](#_ENREF_4)]، [[5](#_ENREF_5)].

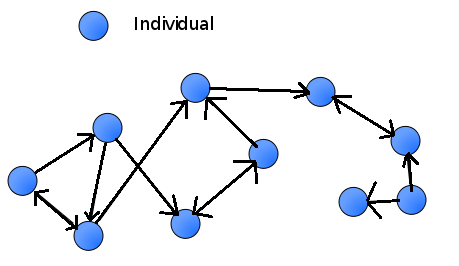
تحلیل شبکه­های اجتماعی به خصوص کشف انجمن در این شبکه­ها از مباحث به روز و پرکاربرد در داده­کاوی است. با توجه به رشد روز افزون این شبکه­ها و گسترش دامنه نفوذ و تاثیرگذاری این شبکه­ها در همه حوضه­ها به ویژه در تجارت، بازاریابی، تبلیغات، سیاست، اخبار، فعالیت­های اجتماعی، هنر و ... امروزه بسیاری از مؤسسات و گروه­های تحقیقاتی به سرمایه­گذاری و تحقیق در این حوزه روی آورده­اند.

تحلیل شبکه­های اجتماعی معمولاً بر اساس نظریه گراف انجام می­شود. به این ترتیب که هر شبکه به صورت یک گراف متشکل از تعدادی گره و یال در نظر گرفته می­شود. در تحلیل شبکه­های اجتماعی ماتریس متناظر این گراف به عنوان یک سری داده معنی­دار مورد پردازش قرار می­گیرد. در مساله کشف انجمن هدف تحلیل اطلاعات این ماتریس به یافتن گروهي از رئوس است که با یالهای زیادی به یكدیگر متصل هستند در حاليكه تعداد یالهای بسيار كمتری، رئوس این گروه را به باقي شبكه متصل مي‎نماید. این مجموعه از رئوس انجمن ناميده مي‎شود.

## 1-2 نظریه گراف در شبکه‎های اجتماعی

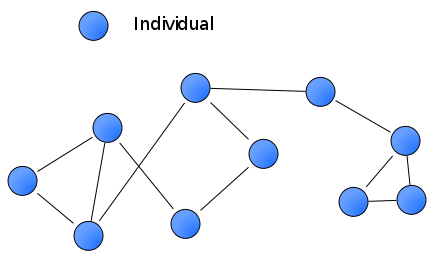
شبکه­های اجتماعی معمولاً روابط بین افراد در ارتباطات روزانه­شان را نشان می­دهند. در این گراف هر راس[[4]](#footnote-4) یا گره[[5]](#footnote-5) متناظر با یکی از افراد عضو آن شبکه است و هر یال[[6]](#footnote-6) یا رابطه[[7]](#footnote-7) ارتباط بین آن دو شخص متناظر با دو گره بین آن یال را نشان می­دهد. مساله کشف انجمن را می­توان به دید مسئله‎ای برای یافتن تجمع رابطه‎ها[[8]](#footnote-8) یا یال‎ها در گراف متناظر شبکه مورد بررسی قرار داد. به این ترتیب که با توجه به رابطه­های موجود بین گره­های یک گراف شبکه سعی شود از روی ازدحامی که در روابط بین گره‎ها وجود دارد، مجموعه‎هایی را پیدا کرد که روابط بین افراد داخل مجموعه نسبت به روابط بین اعضای این مجموعه با اعضای خارج از مجموعه به طرز قابل معنی‎داری بیشتر است. پیدا کردن چنین مجموعه‎هایی در گراف شبکه را کشف انجمن گویند.

انواع زیادی از یال­ها می­توانند بین گره­ها وجود داشته باشند که در ساده­ترین شکل مطالعه شبکه­های اجتماعی نگاشتی از همه یال­های موجود بین گره­ها مورد مطالعه قرار می­گیرد. از چنین گرافی می­توان برای مطالعه موقعیت اجتماعی هر یک از افراد عضو در شبکه استفاده کرد. در برخی از شبکه­های اجتماعی امکان وجود ارتباط یک طرفه وجود دارد. به این ترتیب که شخص «الف» شخص «ب» را در فهرست دوستان خود قرار داده است، اما شخص «ب» درخواست دوستی شخص «ب» را نپذیرفته است و شخص «ب» در فهرست دوستان شخص «الف» قرار ندارد. نیاز است که برای مدل‎سازی چنین شبکه­هایی با گراف متناظر آن از یال­های جهت­دار استفاده شود. اما در بسیاری از شبکه­های اجتماعی مانند فیسبوک و گوگل پلاس امکان برقراری ارتباط یک طرفه وجود ندارد. و برای برقراری یک ارتباط دوستی بین دو نفر در این شبکه­ها تایید این رابطه از طرف هر دو نفر نیاز است. برای چنین شبکه­هایی نیازی به گراف جهت­دار وجود ندارد. در شکل (1-1) یک گراف جهت دار نشان داده شده است.



شکل (1- 1): **نمونه­ای از گراف جهت دار یک شبکه**

جهت پیکان در این شکل نشان دهنده جهت دوستی است. به این معنی که اگر جهت پیکان از شخص «الف» به سمت شخص «ب» باشد، آنگاه شخص «ب» در فهرست دوستان شخص «الف» قرار دارد. در شکل (1-2) نمونه­ای از گراف یک شبکه بدون یال­های جهت دار نشان داده شده است.



شکل (1- 2)**: نمونه­ای از گراف بدون جهت یک شبکه**

## 1-3 بیان مسئله

با نفوذ روزافزون شبکه­های اجتماعی در زندگی و تعاملات مردم بسیاری از گروه­های تحقیقاتی در سراسر جهان علاقه­مند به مطالعه در این زمینه شده­اند. شبکه­ها به صورت گراف مدل می­شود که در این گراف هر گره نماینده یک فرد مشخص و هر یال بین دو گره نشان دهنده­ی وجود تعاملاتی بین این افراد است. افراد در یک شبکه­های اجتماعی با تشکیل انجمن­ها با یکدیگر تبادل اطلاعات می­کنند. آشکارسازی ساختار شبکه یکی از موضوعات اساسی در مطالعه شبکه­های پیچیده است. یکی از نکات قابل بررسی مهم در تحلیل شبکه­ها رفتار پویای آن­ها است، که در واقع تحولات آن­ها را در طول زمان ارزیابی می­کند.

شبکه­های پویا با دنبال کردن تغییرات ساختار شبکه در گام­های زمانی مختلف ارتباطات داخلی خود را بهبود می­بخشند. برای تحلیل و ارزیابی زمانی شبکه­های پویا روش­های مختلفی در مقالات پیشنهاد شده است [[6](#_ENREF_6), [7](#_ENREF_7)] خوشه­بندی تکاملی یکی از روش­هایی است که برخی از مقالات [[7-10](#_ENREF_7)] برای ارزیابی خوشه­ها در داده­های زمانی پیشنهاد داده­اند. این مفهوم برای اولین بار در [[11](#_ENREF_11)] معرفی شد.

داده­های مربوط به گروه­های خوشه­بند تکاملی با معرفی چهارچوبی به نام همواری زمانی[[9]](#footnote-9) برای تشکیل ترتیبی از خوشه­بندی در گام­های زمانی مختلف می­رسند. در این چهاچوب تغییرات ناگهانی خوشه­بندی در یک بازه­ی زمانی کوتاه مطلوب نیست. همواری با برقراری تعادل بین دو شاخص مختلف تحقق می­یابد. اولین شاخص کیفیت لحظه­ای[[10]](#footnote-10) است و بر این منطق استوار است که خوشه­بندی باید بگونه­ای انجام شود که دقت آن تا حد امکان برای داده­های رسیده در گام جاری بالا باشد. دومین شاخص ارزش دوره­ای[[11]](#footnote-11) نام دارد و بر این منطق استوار است که خوشه­بندی نباید به گونه­ای باشد که از یک گام زمانی تا گام زمانی جابجایی چشم­گیری داشته باشد.

در این تحقیق پیشنهاد می­شود برای کشف انجمن در شبکه­ پس از تعریف یک تابع هزینه چندمنظوره که هر دوی این شاخص­ها را دربر بگیرد، از الگوریتم­های جستجوی تکاملی نظیر الگوریتم ژنتیک، رقابت استعماری، تبرید شبیه­سازی شده یا سایر الگوریتم­های تکاملی مشابه استفاده شود.

## 1-4 اهداف و نوآوری‎های پژوهش

این بخش پس از پایان کار با توجه به اهداف محقق شده تکمیل می‎شود.

پرسش اصلی که این پژوهش به دنبال یافتن پاسخی برای آن است می‎تواند به این صورت بیان شود که «آیا می­توان با توسعه روش­های تکاملی چند هدفه کشف انجمن در شبکه­های اجتماعی را دقیق­تر انجام داد؟» برای پاسخگویی به این پرسش یکسری مطالعاتی صورت گرفته است و سپس روشی پیشنهاد شده است که در آن ابتدا شبکه به صورت یک گراف تعریف می­شود. این گراف شامل تعدادی از رئوس به عنوان افراد و یال‎های بین آن­ها برای نشان دادن ارتباط بین افراد است. هر انجمن نیز به صورت یک زیر گراف در آن تعریف می­شود. مسأله کشف انجمن به صورت یک مسأله بهینه سازی تعریف شده است که هدف آن خوشه­بندی افراد در انجمن­هاست. برای این منظور یک تابع هدف چندمنظوره تعریف خواهد شد و برای بهینه­سازی این تابع هدف از الگوریتم­های تکاملی استفاده می‎شود.

اهدافی که در پیشنهاد اولیه این پژوهش مطرح شد عبارست است از:

1. بررسی روش­های کشف انجمن در شبکه­ها و مفاهیم مربوط به آن
2. آشکارسازی ساختار انجمن­ها در شبکه
3. افزایش دقت خوشه­بندی در روش­های کشف انجمن با تعریف تابع هزینه چندمنظوره

برای تحقق این اهداف دو فرضیه زیر در نظر گرفته شده است:

1. در گراف شبکه بین رئوس هر انجمن تعداد زیادی یال وجود دارد و بین رئوس دو انجمن متفاوت یال­های کمتر و پراکنده­تری وجود دارد.
2. شبکه به صورت دینامیک است و ساختار انجمن­ها در گام­های زمانی مختلف ممکن است دچار تغییر شود.

هرچند الگوریتم‎های تکاملی چندمنظوره برای گراف‎های استاتیک پارتیشن بندی شده و برای خوشه بندی داده‎ها پیشنهاد شده است، اما استفاده از آن‎ها برای شبکه‎های پویا به اندازه اهمیت آن مورد بررسی قرار نگرفته است. و می‎توان با بررسی الگوریتم‎های تکاملی مختلف برای حل این مسأله به دقت بالاتری در کشف انجمن رسید.

اهداف تحقق یافته این پژوهش عبارت است از:

1. ................
2. .................
3. ..........................

1-5 ساختار پایان‎نامه

این پایان‎نامه در پنج فصل نگارش شده است. در فصل اول کلیات تحقیق شامل مقدمه، بیان مسئله، ضرورت انجام پژوهش، اهداف و فرضیات است. در این فصل مفاهیم اصلی مانند شبکه‎های اجتماعی، انجمن‎ها و گراف تعریف شده است. سپس مسئله کشف انجمن و راه حل پیشنهادی برای حل این مسئله به صورت کلی بیان شده است.

فصل دوم به بیان مبانی و پیشینه‎ی تحقیق می‎پردازد. این فصل با بیان مبانی نظری و تعریف اصطلاحات و متغیرها شروع می‎شود و سپس به بیان پیشینه‎ی تحقیق و یافته‎های اخیر در این زمینه در خارج و داخل کشور می‎پردازد.

فصل سوم به بیان روش‎ اختصاص دارد. این فصل در دو بخش ارائه شده است در بخش نخست ابتدا روش‎های متداول کشف انجمن در شبکه‎های اجتماعی مرور خواهد شد و در پس از آن روش پیشنهادی در این پژوهش معرفی می‎شود.

در فصل چهارم نتایج حاصل از شبیه‎سازی گزارش شده است. شبیه‎سازی‎هایی انجام شده در این بخش مربوط به روش‎های ارائه شده در فصل سوم می‎باشد.

فصل پنجم آخرین فصل این پایان‎نامه است که در آن پس از بحث و بررسی در مورد نتایج شبیه‎سازی‎های نتیجه‎گیری نهایی انجام شده است و پس از پیشنهاداتی برای ادامه کار و توسعه پژوهش ارائه شده است.

# 2- فصل دوم: پیشینه‎ی تحقیق

## 2-1 مبانی نظری پژوهش

در این بخش به بیان مفاهیم نظری پژوهش پرداخته شده است که در آن برخی تعاریف و مفاهیم اساسی مورد استفاده در پژوهش بیان شده است.

### 2-1-1 داده‎کاوی

داده کاوی عبارت است از فرآيند (نيمه) خودکار استخراج دانش (در قالب الگوهای پنهان) از مجموعه اطلاعات ورودی. وظیفه ی داده‎کاوی، کاویدن و استخراج از منابع عظیم داده است تا اطلاعات گرانبهایی که در حجم انبوهی از اطلاعات سطحی پنهان شده است را استخراج کند. معمولاً آگاهی اندکی در مورد دانش هدف وجود دارد و ورودی عمدتا بسيار حجيم و پردازش دستی آن ناممکن است. نتايج حاصل از داده کاوی، با روش‎های سنتی پردازش اطلاعات (گزارش گيری) قابل دستيابی نيست و خودکار يا نيمه خودکار بودن داده کاوی به معنای حداقل نياز به دخالت کاربر است. در داده‎کاوی انواع اطلاعات (و نه صرفاً اطلاعات عددی) قابل پردازش می‎باشند. داده‎کاوی را می‎توان يک شاخه از يادگيری ماشين دانست. به دليل عدم وجود يک چارچوب تئوريک برای داده‎کاوی، در نظر گرفتن آن به عنوان زير مجموعه‎ای از يادگيری ماشين می‎تواند مورد بحث قرار گيرد. تفاوت اصلی داده کاوی و علم آمار، در حجم داده های مورد تحلیل، روش مدلسازی داده ها و استفاده از هوش مصنوعی است.

تلاش‎های اندکی برای توسعه يک چارچوب نظری برای داده‎کاوی انجام گرفته است. اين ميزان تلاش کافی نبوده و به نتيجه قابل قبولی هم نرسيده است. تفسير داده‎کاوی به عنوان زير مجموعه‎ای از آمار، چندان قانع کننده نيست. به عنوان مثال مسائل با فضای حالت دارای ابعاد زياد مهمترين وجه اين تمايزند. تفسير داده‎کاوی به عنوان فرآيندی جهت تخمين تابع توزيع احتمال توأم نمونه‎ها نیز چندان مناسب نیست، به عنوان مثال تکراری بودن داده کاوی چنين طبقه بندی را رد می کند.

تعاريف موجود، سبب ايجاد انتظاراتی غيرواقعی از داده کاوی می‎شود. تا کنون، هيچ سيستم/فرآيند داده کاوی کاملاً خودکاری که منطبق بر شرايط واقعی باشد، ساخته نشده است. دخالت مستقيم کاربر، به خصوص در مراحل اوليه يک فرآيند داده کاوی، اجتناب ناپذير است. داده‎کاوی به يک هيولای پرقدرت شبيه است: رها کردن بدون هدف آن در سرزمين اطلاعات، مطمئناً نتايج خوبی در پی نخواهد داشت.

### 2-1-2 انتخاب ویژگی

مسئله انتخاب ویژگی، یکی از مسائلی است که در مبحث یادگیری ماشین[[12]](#footnote-12) و همچنین شناسائی الگو مطرح است. این مسئله در بسیاری از کاربردها مانند طبقه‌بندی اهمیت به سزائی دارد، زیرا در این کاربردها تعداد زیادی ویژگی وجود دارد که بسیاری از آنها یا بلااستفاده هستند و یا اینکه بار اطلاعاتی چندانی ندارند [[12](#_ENREF_12)].

حذف نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد مورد نظر بالا می‌برد و علاوه بر این باعث می شود که اطلاعات غیر مفید زیادی را به همراه داده‌های مفید ذخیره کنیم [[13](#_ENREF_13)].

همچنین ممکن است ویژگی‌هایی وجود داشته باشند که بار اطلاعاتی منفی داشته باشند به این معنی که استفاده از این ویژگی‌ها نه تنها کمکی به بهبود کارآیی طبقه‌بند نمی‌کند، بلکه منجر به گمراه شدن طبقه‌بند شده و نتیجه به دست آمده را بدتر می‌کنند. از این رو لازم است برای افزایش دقت طبقه‌بند، این ویژگی‌ها را یافته و از مجموعه ویژگی‌ها حذف کرد [[14](#_ENREF_14)].

برای مسئله انتخاب ویژگی، راه حل ها و الگوریتم‌های فراوانی ارائه شده است که بعضی از آنها قدمت سی یا چهل ساله دارند. مشکل بعضی از الگوریتم ها در زمانی که ارائه شده بودند، بار محاسباتی زیاد آنها بود، اگر چه امروزه با ظهور کامپیوترهای سریع و منابع ذخیره سازی بزرگ این مشکل، کمتر به چشم می‌آید ولی از طرف دیگر، مجموعه‌های داده‌ای بسیار بزرگ برای مسائل جدید باعث شده است که همچنان پیدا کردن یک الگوریتم سریع برای این کار مهم باشد [[15](#_ENREF_15)].

در این بخش ما در ابتدا تعاریفی که برای انتخاب ویژگی ارائه شده‌اند و همچنین، تعاریف مورد نیاز برای درک این مسئله را ارائه می دهیم. سپس روش های مختلف برای این مسئله را بر اساس نوع و ترتیب تولید زیرمجموعه ویژگی‌های کاندید و همچنین نحوه ارزیابی این زیرمجموعه ها دسته بندی می‌کنیم. در نهایت تعدادی از روش‌های معرفی شده را معرفی و بر اساس اهمیت، تا جائی که مقدور باشد، آنها را تشریح و الگوریتم برخی از آنها را ذکر می‌کنیم.

مسئله انتخاب ویژگی بوسیله نویسندگان مختلف، از دیدگاه‌های متفاوتی مورد بررسی قرار گرفته و هر نویسنده نیز با توجه به نوع کاربرد، تعریفی را از آن ارائه داده است.

در ادامه چند مورد از این تعاریف را بیان می کنیم:

**1. تعریف ایده آل:** پیدا کردن یک زیرمجموعه با حداقل اندازه ممکن، برای ویژگی‌ها است، که برای هدف مورد نظر اطلاعات لازم و کافی را در بر داشته باشد. بدیهی است که هدف تمام الگوریتم‌ها و روش‌های انتخاب ویژگی، یافتن همین زیر مجموعه است.

**2. تعریف کلاسیک:** انتخاب یک زیرمجموعه M عنصری از میان N ویژگی، به طوریکه M < N باشد و همچنین مقدار یک تابع ارزیابی برای زیرمجموعه مورد نظر، نسبت به سایر زیرمجموعه‌های هم اندازه دیگر بهینه باشد.

**3. افزایش دقت پیشگوئی:** هدف انتخاب ویژگی این است که یک زیرمجموعه از ویژگی‌ها برای افزایش دقت پیشگوئی انتخاب شوند. به عبارت دیگر کاهش اندازه ساختار بدون کاهش قابل ملاحظه در دقت پیشگوئی طبقه‌بندی کننده ای که با استفاده از ویژگی‌های داده شده به دست می آید.

**4. تخمین توزیع کلاس اصلی:** هدف از انتخاب ویژگی این است که یک زیرمجموعه کوچک از ویژگی‌ها انتخاب شوند، توزیع ویژگی‌هایی که انتخاب می شوند، بایستی تا حد امکان به توزیع کلاس اصلی با توجه به تمام مقادیر ویژگی‌های انتخاب شده نزدیک باشد [[17](#_ENREF_17)].

روش‌های مختلف انتخاب ویژگی، تلاش می کنند تا از میان 2n زیر مجموعه کاندید، بهترین زیرمجموعه را پیدا کنند. در تمام این روش‌ها بر اساس کاربرد و نوع تعریف، زیر مجموعه ای به عنوان جواب انتخاب می شود که بتواند مقدار یک تابع ارزیابی را بهینه کند. با وجود این که هر روشی سعی می‌کند بهترین ویژگی‌ها را انتخاب کند، اما با توجه به وسعت جواب‌های ممکن، پیدا کردن جواب بهینه مشکل و در N‌ های متوسط و بزرگ بسیار پرهزینه است [[18](#_ENREF_18)].

در حالت کلی می توان فرایند انتخاب ویژگی را به چهار قسمت تقسیم کرد:

**1. تابع تولید کننده:** این تابع زیر مجموعه‌های کاندید را معین می کند.

**2. تابع ارزیابی:** زیرمجموعه مورد نظر را طبق روشی، ارزیابی کرده و یک عدد را به عنوان میزان خوبی روش باز می‌گرداند.

**3. شرط خاتمه:** زمان توقف الگوریتم را مشخص می کند.

**4. تابع تعیین اعتبار:** تصمیم می‌گیرد که آیا زیر مجموعه انتخاب شده معتبر است یا خیر؟

تابع تولید کننده در واقع تابع جستجو است. این تابع زیرمجموعه­های مختلف را به ترتیب تولید می­کند تا به وسیله تابع ارزیابی، مورد ارزیابی قرا بگیرد. تابع تولید کننده از یکی از حالت­های زیر شروع به کار می­کند:

1. بدون ویژگی
2. با مجموعه تمام ویژگی­ها
3. با یک زیرمجموعه تصادفی

در حالت اول ویژگی­ها به ترتیب به مجموعه اضافه می­شوند و زیرمجموعه­های جدید را تولید می­کنند. این عمل آنقدر تکرار می­شود تا به زیرمجموعه مورد نظر برسیم. به اینگونه روش­ها، روش‌های پایین به بالا می‌گویند.

در حالت دوم از یک مجموعه شامل تمام ویژگی­ها، شروع می­کنیم و به مرور و در طی اجرای الگوریتم، ویژگی­هایی را حذف می­کنیم تا به زیرمجموعه دلخواه برسیم. روش­هایی که به این صورت عمل می‌کنند، روش‌های بالا به پایین نام دارند.

یک تابع ارزیابی، میزان خوب بودن یک زیرمجموعه تولید شده را بررسی کرده و یک مقدار به عنوان میزان خوب بودن زیرمجموعه مورد نظر بازمی­گرداند. این مقدار با بهترین زیرمجموعه قبلی مقایسه می‌شود. اگر زیرمجموعه جدید، بهتر از زیرمجموعه­های قدیمی باشد، زیرمجموعه جدید به عنوان زیرمجموعه بهینه، جایگزین قبلی می­شود.

باید توجه داشت که بدون داشتن یک شرط خاتمه مناسب، فرآیند انتخاب ویژگی ممکن است، برای همیشه درون فضای جستجو، برای یافتن جواب سرگردان بماند. شرط خاتمه می­تواند بر پایه تابع تولید کننده باشد، مانند:

1. هر زمان که تعداد مشخصی ویژگی انتخاب شدند.
2. هر زمان که به تعداد مشخصی تکرار رسیدیم.

و یا اینکه بر اساس تابع ارزیابی انتخاب شود، مانند:

1. وقتی که اضافه یا حذف کردن ویژگی، زیرمجموعه بهتری را تولید نکند.
2. وقتی که به یک زیرمجموعه بهینه بر اساس تابع ارزیابی برسیم.

تابع تعیین اعتبار جزئی از فرآیند انتخاب ویژگی نیست، اما در عمل بایستی یک زیرمجموعه ویژگی را در شرایط مختلف امتحان کنیم تا ببینیم که آیا شرایط مورد نیاز برای حل مسئله مورد نظر ما را دارد یا نه. برای این کار می‌توانیم از داده­های نمونه­برداری شده و یا مجموعه داده‌های شبیه سازی شده استفاده کنیم.

### 2-1-3 طبقه‌بندی روش­های مختلف انتخاب ویژگی

روش­های مختلف انتخاب ویژگی را می‌توان بر اساس دو معیار تابع تولید کننده و تابع ارزیابی طبقه‌بندی کرد. در ادامه این روش ها را مورد بررسی قرار می‌دهیم.

ابتدا از نظر تابع تولید کننده این روش‎ها مورد ارزیابی قرار می‎گرد. اگر تعداد کل ویژگی­ها برابر N باشد، تعداد کل زیرمجموعه­های ممکن برابر 2N می­شود. این تعداد برای N‌های متوسط هم خیلی زیاد است. بر اساس نحوه جستجو در میان این تعداد زیر مجموعه، روش‌های مختلف انتخاب ویژگی را می­توان به سه دسته زیر تقسیم­بندی نمود:

1. جستجوی کامل
2. جستجوی مکاشفه­ای
3. جستجوی تصادفی

در ادامه به معرفی هر کدام از این دسته­ها می­پردازیم.

1. **جستجوی کامل:** در روش­هایی که از این نوع جستجو استفاده می­کنند، تابع تولید کننده بر اساس تابع ارزیابی استفاده شده، تمام فضای جواب (زیرمجموعه­های ممکن) را برای یافتن جواب بهینه جستجو می‌کند. توابع مکاشفه­ای مختلف زیادی طراحی شده­اند، تا جستجو را بدون از دست دادن شانس پیدا کردن جواب بهینه، کاهش دهند. اما با توجه به بزرگی فضای جستجو، O(2N)، این روش­ها باعث می­شوند که فضای کمتری جستجو شود. روش­ها و تکنیک­های مختلفی برای این کار استفاده شده‌اند و بعضی از آنها از تکنیک بازگشت به عقب نیز در جریان کار استفاده کرده­اند [[19](#_ENREF_19)].
2. **جستجوی مکاشفه­ای:** در روش­های با این نوع جستجو، در هر بار اجرای الگوریتم، یک ویژگی به مجموعه ویژگی انتخاب شده، اضافه و یا از آن حذف می­شود. به همین دلیل پیچیدگی زمانی آنها محدود و کمتر از O(N2) می‌باشد. در اینگونه موارد، اجرای الگوریتم نسبتا سریع می­باشد و پیاده‌سازی آنها نیز بسیار ساده است.
3. **جستجوی تصادفی:** روش­هایی که از این نوع جستجو استفاده می­کنند، محدوده کمتری از فضای کل حالات را جستجو می‌کنند، که اندازه این محدوده به حداکثر تعداد تکرار الگوریتم بستگی دارد. در این روش­ها پیدا شدن جواب بهینه به اندازه منابع موجود و زمان اجرای الگوریتم بستگی دارد. در هر بار تکرار، تابع تولید کننده تعدادی از زیرمجموعه­های ممکن از فضای جستجو را به صورت تصادفی انتخاب می­کند و در اختیار تابع ارزیابی قرار می­دهد. تابع تولید کننده تصادفی پارامتر­هایی دارد که بایستی تنظیم شوند، تنظیم مناسب این پارامترها در سرعت رسیدن به جواب و پیدا شدن جواب­های بهتر موثر است. این الگوریتم‌ها را با نام الگوریتم‌های فرامکاشفه‌ای نیز می‌شناسند.

در ادامه انتخاب ویژگی با توجه به تابع ارزیابی مورد بررسی قرار گرفته است. پیدا شدن یک زیرمجموعه بهینه از مجموعه ویژگی­ها، به صورت مستقیم به انتخاب تابع ارزیابی بستگی دارد. چرا که اگر تابع ارزیابی به زیرمجموعه ویژگی بهینه یک مقدار نامناسب نسبت دهد، این زیرمجموعه هیچ‌گاه به عنوان زیرمجموعه بهینه انتخاب نمی­شود. مقادیری که توابع ارزیابی مختلف به یک زیرمجموعه می­دهند، با هم متفاوت است.

توابع ارزیابی را می­توان به طرق مختلفی دسته بندی کرد. در اینجا ما دسته‌بندی­ای را که توسط Dash و Liu ارائه شده است، بیان می­کنیم. آنها این معیارها را به پنج دسته تقسیم کرده­اند:

1. **معیارهای مبتنی بر فاصله[[13]](#footnote-13) :** در این معیارها، مثلاً برای یک مسئله دو کلاسه، یک ویژگی یا یک مجموعه ویژگی مثل X بر یک ویژگی یا یک مجموعه ویژگی دیگر مثل Y ارجحیت دارد، اگر که با آن مجموعه ویژگی مقادیر بزرگتری برای اختلاف بین احتمالات شرطی دو کلاس داشته باشیم. نمونه­ای از این معیارها همان معیار فاصله اقلیدسی می­باشد.
2. **معیارهای مبتنی بر اطلاعات**[[14]](#footnote-14): این معیارها میزان اطلاعاتی را که بوسیله یک ویژگی به دست می­آید را در نظر می­گیرند. ویژگی X در این روش­ها بر ویژگی Y اولویت دارد، اگر اطلاعات بدست آمده از ویژگی X بیشتر از اطلاعاتی باشد، که از ویژگی Y به دست می­آید. نمونه‌ای از این معیارها، معیار آنتروپی می­باشد.
3. **معیارهای مبتنی بر وابستگی**[[15]](#footnote-15): این معیارها که با عنوان معیارهای همبستگی[[16]](#footnote-16) نیز شناخته می‌شوند، قابلیت پیشگوئی مقدار یک متغیر بوسیله یک متغیر دیگر را اندازه­گیری می­کنند. ضریب تاثیر[[17]](#footnote-17) یکی از معیارهای وابستگی کلاسیک است و می­توانیم آن را برای یافتن همبستگی بین یک ویژگی و یک کلاس به کار ببریم. اگر همبستگی ویژگی X با کلاس C بیشتر از همبستگی ویژگی Y با کلاس C باشد، در این‌صورت ویژگی X بر ویژگی Y برتری دارد. با یک تغییر کوچک، می­توانیم وابستگی یک ویژگی با ویژگی­های دیگر را اندازه­گیری کنیم. این مقدار درجه افزونگی این ویژگی را نشان می­دهد.
4. **معیارهای مبتنی بر سازگاری**[[18]](#footnote-18): این معیارها جدیدتر هستند و اخیراً توجه بیشتری به آنها شده است. این معیارها خصوصیات متفاوتی نسبت به سایر معیارها دارند، زیرا که به شدت به داده­های آموزشی تکیه دارند و در انتخاب یک زیرمجموعه از ویژگی­ها تمایل دارند که مجموعه ویژگی­های کوچکتری را انتخاب کنند. این روش­ها زیرمجموعه­های با کمترین اندازه را بر اساس از دست دادن یک مقدار قابل قبول سازگاری که توسط کاربر تعیین می­شود، پیدا می­کنند.
5. **معیارهای مبتنی بر خطای طبقه‌بندی کننده**[[19]](#footnote-19): روش­هایی که این نوع از تابع ارزیابی را استفاده می‌کنند، با عنوان "wrapper methods" شناخته می­شوند. دقت عملکرد در این روش­ها برای تعیین کلاسی که نمونه داده شده متعلق به آن است، برای نمونه­های دیده نشده بسیار بالا است، اما هزینه‌های محاسباتی در آنها نیز نسبتاً زیاد است و زمان طولانی‌تری مصرف می‌کنند.

در جدول 2-1 مقایسه­ای بین انواع مختلف تابع ارزیابی، صرف نظر از نوع تابع تولید کننده مورد استفاده، انجام شده است.

جدول (1- 1) : انواع توابع ارزیابی و ویژگی‌های آنها

|  |  |  |  |
| --- | --- | --- | --- |
| **دقت** | **پیچیدگی زمانی** | **عمومیت** | **نوع تابع ارزیابی** |
| **-** | پایین | دارد | معیار فاصله |
| **-** | پایین | دارد | معیار اطلاعات |
| **-** | پایین | دارد | معیار وابستگی |
| **-** | متوسط | دارد | معیار سازگاری |
| **خیلی زیاد** | بالا | ندارد | خطای طبقه‌بندی کننده |

پارامترهایی که برای مقایسه استفاده شده­اند به صورت زیر می­باشند:

1. **عمومیت[[20]](#footnote-20)**: اینکه بتوان زیرمجموعه انتخاب شده را برای طبقه­بندی کننده­های متفاوت به کار ببریم.
2. **پیچیدگی زمانی**: زمان لازم برای پیدا کردن زیرمجموعه ویژگی­ جواب.
3. **دقت**: دقت پیشگوئی با استفاده از زیرمجموعه انتخاب شده.

علامت "-" که در ستون آخر آمده است، به این معنی است که در مورد میزان دقت حاصل نمی‌توانیم مطلبی بگوئیم. به جز خطای طبقه‌بندی کننده، دقت سایر توابع ارزیابی به مجموعه داده مورد استفاده و طبقه‌بندی کننده‌ای که بعد از انتخاب ویژگی برای طبقه‌بندی کلاس‌ها استفاده می‌شود، بستگی دارد. این توابع در زمره روش‌های فیلتر محسوب می‌شوند که ساده بوده و پیچیدگی محاسباتی کمی دارند، در نتیجه از سرعت نسبتا بالایی برخوردارند.

### 2-1-4 یادگیری ماشین و طبقه­بندی

*یادگیری ماشین[[21]](#footnote-21) شاخه مهمی از گرایش هوش مصنوعی است که هدف آن تعلیم یک ماشین یادگیرنده برای ایجاد یک مدل پیش بینی است، به نحوی که بر اساس آن، ماشین بتواند از طریق تجربه یادگیری کرده و عملکرد خود را بهتر کند[*[*20*](#_ENREF_20)*]. علم یادگیری ماشین به صورت سنتی به سه شاخه اصلی تقسیم می­شود: 1. یادگیری بدون ناظر[[22]](#footnote-22) 2.یادگیری همراه با ناظر[[23]](#footnote-23) 3.یادگیری تقویتی.*

*در یادگیری ماشین، هدف آموزش ماشین برای ایجاد یک مدل پیش بینی بر اساس تجربه می باشد. در برخی مسائل داده های بسیاری جمع آوری گشته شده است، اطلاعات پیرامون داده های موجود به اندازه کافی بوده و هدف طراحی ماشینی است که بتواند به صورت هوشمندانه داده­های جدیدی را که به دست می­آید، طبقه­بندی نماید. در برخی مسائل مانند مسائل زیستی به علت ناشناخته بودن عوامل محیطی مسئله، حجم بالای داده­های به دست آمده پیرامون مسئله و از آنجا حجم اطلاعات به دست آمده، نیاز به طراحی ماشینی هست که به صورت هوشمندانه الگوهای موجود در داده­ها را استخراج نموده و بتواند داده­ها را خوشه­بندی[[24]](#footnote-24) نماید. از این رو معمولا یادگیری ماشین را در سه بخش اصلی بررسی می نمایند که عبارتند از:*

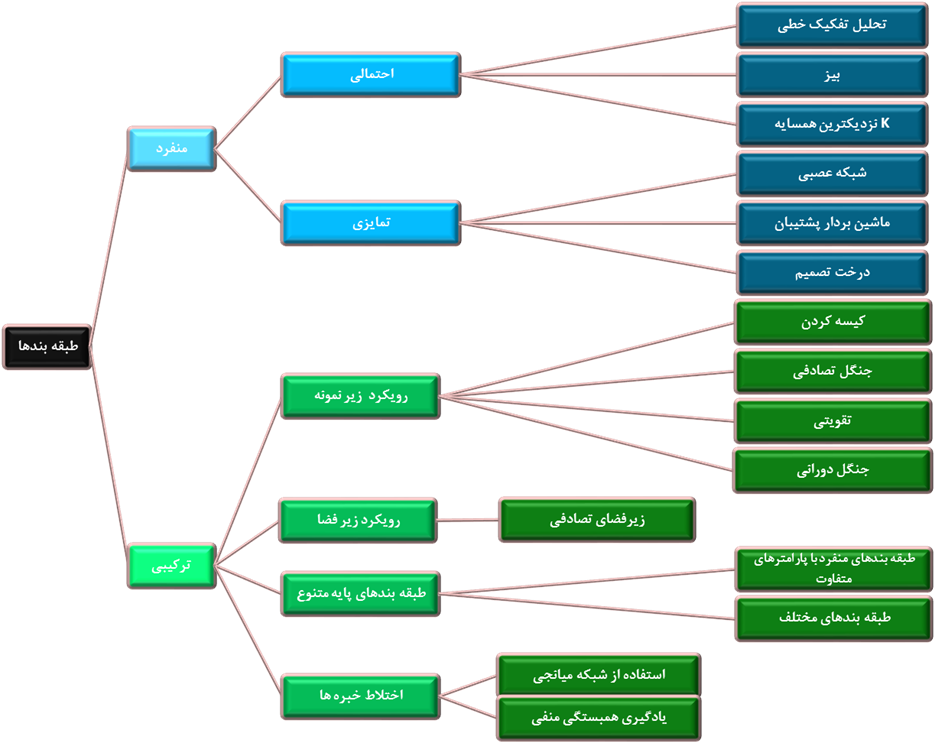
*1. یادگیری بدون ناظر: در این روش یادگیری، مجموعه­ای از داده­های بدون برچسب[[25]](#footnote-25) توسط مدل مشاهده می­شود. هدف از این روش سامان دادن به داده­ها یا به عبارت دیگر یافتن نظم خاصی در میان داده­هاست. خوشه­بندی[[26]](#footnote-26) داده­ها، تشخیص داده­های پرت و کاهش بعد داده­ها، روش­ها و کاربردهایی از یادگیری بدون ناظر هستند.*

*2. یادگیری همراه با ناظر: در این روش یادگیری، مجموعه­ای از داده­های با برچسب توسط مدل مشاهده می­شود. هدف این روش، یافتن قوانین حاکم بر داده­ها و روابط نهفته میان آن­هاست، به­طوری­که پس از آموزش مدل قادر به تشخیص برچسب داده­های مشاهده ­نشده نیز باشد. زماني­كه برچسب داده­ها، عضو مجموعه اعداد حقيقي باشند، يادگيري همراه با ناظر رگرسيون[[27]](#footnote-27) و زماني­كه بر چسب داده­ها گسسته باشد، يادگيري همراه با ناظر، طبقه­بندی[[28]](#footnote-28) نامیده می­شود.*

*3. یادگیری تقویتی: در اين روش يادگيري، مدل بصورت تكراري داده­ها را مشاهده كرده و عمل خاصي را بر روي يك داده انجام مي­دهد و در قبال عمل انجام شده پاداش معيني دريافت مي­كند. هدف نهايي اين مدل، انتخاب اعمالي است كه در آينده پاداش بيشينه را دريافت كند.*

*در ادامه ایزار یادگیری ماشین در حوزه یادگیری ماشین همراه با ناظر که تحت عنوان طبقه­بند[[29]](#footnote-29)ها آن­ها را می­شناسند، معرفی خواهند گشت.*

*در این بخش، ابزار یادگیری ماشین که امروزه در بسیاری از زمینه های تحقیقاتی نظیر پردازش تصویر، داده­کاوی و بیوانفورماتیک استفاده کاربردی دارند، به صورت مشروح بیان خواهند شد. طبقه­بندها را می­توان در دو گروه اصلی دسته­بندی کرد. طبقه­بند­های منفرد و طبقه­بندهای ترکیبی . این دسته­بندی در شکل شماره (2-3) نمایش داده شده است. ابتدا طبقه­بندهای منفرد شرح داده می­شود و در ادامه تشریح طبقه­بندهای ترکیبی ارائه می­گردد.*

**

شکل (2- 1): دسته بندی کلی انواع طبقه­بندها

*برای طبقه­بندهای منفرد دسته­بندی­های مختلفی وجود دارد که در این پژوهش بر اساس دسته­بندی زیر معرفی خواهند شد[[30]](#footnote-30).*

*فرض کنیم مجموعه داده(نمونه ها) X و مجموعه طبقه(کلاس، برچسب) متناظر با هر یک از داده­ها، یعنی Y در اختیار است و هدف برچسب­دهی این نمونه­های داده،­ از مجموعه طبقه Y می­باشد. در دسته طبقه­بندهای احتمالی، به دنبال یافتن توزیع توأم برای دو متغیر X و Y یعنی P(X,Y) هستند؛ در حالی که در دسته طبقه­بندهای تمایزی، هدف یافتن توزیع شرطی P(Y|X) می­باشد. از طبقه­بندهای مهم در دسته طبقه­بندهای احتمالی می‎توان به طبقه بند بیز و K نزدیکترین همسایه اشاره داشت. در دسته طبقه­بندهای تمایزی، شبکه­های عصبی مصنوعی، ماشین بردار پشتیبان و درختان تصمیم جزء عناصر مهم این دسته به شمار می­روند.*

*در این بخش دو طبقه بند بیز و نزدیکترین همسایه که از طبقه بندهای مهم در دسته طبقه­بندهای احتمالاتی به حساب می آیند، تشریح خواهند شد.*

### 2-1-5 الگوریتم‎های فرامکاشفه‎ای زیستی

در بسياری از مسائل جستجو در دنيای واقعي، نمي‎توان فضای جستجو را به شكل یک درخت نمایش داد. علي الخصوص هنگاميكه این فضا بسيار بزرگ و نامنظم باشد. برای همين برای این نوع فضاها نمي‎توان از هر الگوریتمي استفاده كرد. جستجوی فرامكاشفه‎ای[[31]](#footnote-31) نام خانواده‎ای از الگوریتم‎های جستجوی آگاهانه[[32]](#footnote-32) است كه در آن‎ها از یک پدیده طبيعي برای كاوش فضای جستجو، الهام گرفته مي‎شود. این الگوریتم‎ها با الهام گرفتن از یک رخداد طبيعي این امكان را پيدا مي‎كنند كه فضای جستجوی بسيار بزرگ طيف وسيعي از مسائل بهينه‎سازی‎های پيچيده را به صورت بسيار هوشمندانه‎ای مورد كاوش قرار دهند. منظور از هوشمندانه در اینجا این است كه الگوریتم‎های فرامكاشفه‎ای كل فضای جستجو را به دليل بزرگي آن پيمایش نمي‎كنند. به همين دليل هم هست كه آن‎ها ناكامل و البته غير بهينه هستند. بلكه آنها تنها به پيمایش بخشي از فضا كه احتمال وجود یک پاسخ به اندازه‎ی كافي خوب در آن بيشتر است، اكتفا مي‎كنند. برخي از الهام‎های طبيعي كه بر اساس آن‎ها یک الگوریتم فرامكاشفه‎ای طراحي شده است عبارتند از: تكامل موجودات در طي نسل‎ها، فرآیند سرد شدن یا تبرید در فلزات، زندگي مورچه‎ها در یک كلوني، حركت گروهي پرندگان و سيستم ایمني در بدن انسان. در ادامه معرفي مفاهیمي چون دورنمای برازش، قابليت‎های پویش و انتفاع و نيز دسته بندی الگوریتم‎های فرامكاشفه‎ای خواهيم پرداخت.

در ادامه مختصری در باب طبق‎ بندی الگوریتم‎های فرامكاشفه‎ای از دیدگاه‎های متفاوت و با معيارهای مختلف خواهيم پرداخت. این دیدگاه‎ها عبارتند از:

* **زیستی و غیرزیستی:** بسياری از الگوریتم‎های فرامكاشفه‎ای از زندگي موجودات طبيعي الهام گرفته شده است و به عبارت دیگر زیستي هستند درحاليكه برخي اینگونه نيستند مثلاً براساس یک پدیده فيزیكي الهام گرفته شده‎اند.
* **جمعیتی و غیرجمعیتی: ا**لگوریتم‎های جمعيتي در حين جستجو، یک جمعيت از جواب‎ها را در نظر مي‎گيرند در حاليكه الگوریتم‎های غيرجمعيتي)مبتني بر یک جواب) حين جستجو یک جواب را تغيير مي‎دهند.
* **تکاملی و غیرتکاملی:** الگوریتم‎های زیستي كه به طور مستقيم از فرامكاشفه بقاء اصلح كه مبتني بر نظریه داروین است استفاده مي‎كنند، تكاملي محسوب مي‎شوند. درحاليكه سایر روش‎های جستجوی زیستي به صورت غير مستقيم به فرامكاشفه بقاءاصلح داروین مرتبط هستند زیرا عملاً مولود فرآیند تكامل هستند كه به آن‎ها زیستي غيرتكاملي گفته مي‎شود.
* **باحافظه و بدون حافظه:** برخي الگوریتم‎ها فاقد حافظه‎اند، بدین معنا كه آن‎ها از اطلاعات بدست آمده در طول فرآیند جستجو، استفاده نمي‎كنند )مثل تبرید شبیه‎سازی شده). اما برخي دیگر از حافظه استفاده مي‎كنند.
* **احتمالی و قطعی:** برخي الگوریتم‎های فرامكاشفه‎ای با برخي قوانين احتمالي در حين جستجو دست به تصميم‎گيری مي‎زند، در حاليكه برخي دیگر با تصميمات قطعي مسئله را پيش مي‎برند.

## 2-2 مروری بر پژوهش‎های پیشین

تحلیل شبکه­ها و ارزیابی آن­ها در سال­های گذشته یکی از زمینه­های تحقیقاتی مورد علاقه بسیاری از گروه­های تحقیاتی بوده است و با روند رو به رشدی همراه است [[6-8](#_ENREF_6)]. ساختار شبکه­های اجتماعی معیار خوبی برای پیشبینی عملکرد کاربران است. یکی از مشکلات اساسی در ارتباط با جستجوی روابط بین گره­ها در شبکه در سال­های اخیر جستجوی انجمن به صورت خودکار بوده است [[21](#_ENREF_21)]. علم کشف انجمن در شبکه­ها بسیار غنی است. چندین مورد از الگوریتم­های کشف انجمن در شبکه­های اجتماعی را می­توان برشمرد. همچنین چندین روش برای ارزیابی کیفیت شبکه­ها معین شده است. برای مرور جامع این الگوریتم­ها توصیه می­شود به لسکوک و همکاران در [[9](#_ENREF_9), [22](#_ENREF_22)]، نیومن در [[23](#_ENREF_23)] و دانون و همکاران در [[24](#_ENREF_24)] مراجعه شود.

بسیاری از الگوریتم­های آشکارسازی انجمن از ارتباط بین گره­ها به منظور تعیین چگالی نواحی در گراف استفاده می­کنند [[11](#_ENREF_11)] . این نواحی چگال انجمن­های موجود در شبکه­های اجتماعی در گراف هستند. زو و همکاران در [[10](#_ENREF_10)] الگوریتم جستجوی کشف انجمن جدیدی بر مبنای مدل آماری به منظور سازماندهی کاربران دارای علایق مشابه در یک گروه پیشنهاد کردند. در این مقاله ابتدا مسئله کشف انجمن تشریح شده است و سپس گراف شبکه ترسیم شد. الگوریتم پیشنهادی آن­ها بر این فرض استوار بود که کاربران عضو یک انجمن ممکن است دارای علایق مشابهی باشند. نوآوری اصلی این مقاله این بود که موضوعات انجمن­ها به صورت توزیع چندگانه روی کلمات ارائه می­شود، و علایق کاربران در موضوعات مختلف از یک توزیع آماری از موضوعات انجمن­ها تبعیت می­کند.

جلالی و موسوی در [[25](#_ENREF_25)] به جای استفاده از ویژگی­های آماری متداول گراف شبکه برای شناسایی انجمن در شبکه­های اجتماعی از رفتار و عملکردهای کاربران برای شناسایی گره­های راهنما استفاده کردند و سپس گره­های مشابه آن­ها را مورد شناسایی قرار دادند. هدف آن­ها این بود که با استفاده از ویژگی­ها و فعالیت­هایی که توسط کاربران در شبکه­های اجتماعی برخط انجام می­شود ساختار انجمن­ها را شناسایی کنند.

برودکا و همکاران در [[26](#_ENREF_26)] روشی جدید برای کشف ارزیابی گروه به نام GED پیشنهاد کردند. نتایج آزمایش­ها روی شبکه­ اجتماعی مبتنی بر ایمیل با دو روش ارزیابی دیگر مقایسه شده است و در این مقاله ارائه شده است.

کمار و همکاران [[27](#_ENREF_27)] ارزیابی ویژگی­های شبکه را در دو شبکه بزرگ مورد مطالعه قرار دادند و سعی کردند اعضای هر شبکه را در گروه­هایی خوشه­بندی کنند. سان و همکاران [[28](#_ENREF_28)] روشی بسیار کارآمد و بدون نیاز از پارامتر به نام گراف اسکوپ را معرفی کدند. این روش بر اساس کمترین طول توصیف به منظور جستجوی انجمن در گراف­های شبکه عمل می­کند. اسور و همکاران [[29](#_ENREF_29)] با تعریف رخداد­های مهمی که در شبکه اتفاق میفتد ارزیابی انجمن­ها را مشخص کردند. تانگ و همکاران [[30](#_ENREF_30), [31](#_ENREF_31)] شبکه­های چندگانه را تعریف کردند و یک خوشه­بندی کننده طیفی برای جستجوی انجمن و یافتن ارتباط آن­ها معرفی نمود. چندین روش دیگر نیز برای یافتن انجمن­ها در شبکه­های اجتماعی وجو دارد که در اینجا ذکر نشده است. برای مرو همه آن­ها می­توان به مقاله مروری [[31](#_ENREF_31)] مراجعه کرد. بسیاری از روش­های پیشنهادی از یک تحلیل مرحله­ای شبکه شروع می­شوند. به عنوان مثال ابتدا انجمن­ها استخراج می­شود و سپس تفاوت­های ساختاری آن­ها در طول زمان کشف می­شود تا ارتباطات متعلق بهه شبکه در دو گام زمانی مختلف معین شود.

یک روش­ متفاوت که با نام خوشه­بندی تکاملی شناخته شده است توسط چاکرابرتی و همکاران در [[32](#_ENREF_32)] معرفی شده است. از آنجا در این مقاله توسعه روش­های مبتنی بر خوشه­بندی تکاملی پشنهاد شده است از این رو بیشتر مورد بررسی قرار گرفته است. چاکرابتی و همکاران مشاهده کردند که تغییر ارتباطات در زمان کوتاه ممکن است به علت نویز ایجاد شده باشد. از اینرو گرجه خوشه­بندی در اصل باید با توجه به ارتباطات هر لحظه باشد، اما در بسیاری از حوزه­های کاربردیبهتر است تغییر ناگهانی نسبت به گذشته مورد توجه قرار نگیرد. در هر گام زمانی عمل خوشه بندی انجمن­ها باید با توجه به دو اصل مهم انجام شود. اول اینکه خوشه­بندی باید بازگو کننده ارتباطات موجود در شبکه در هر لحظه (گام زمانی) باشد. دوم اینکه پس از هر طبقه بندی تغییر ناگهانی شدید نسبت به گام­های قبلی نزدیک مشاهده نشود. برای این منظور یک تابع هزینه مناسب تعیف می­شود که هردو خواسته را دربربگیرد و علاوه بر آن قابل حل با روش­های تکاملی باشد.

کیم و همکاران در [[33](#_ENREF_33)] روش مبتنی بر چگالی و ذره پیشنهاد کردند که بر مبنای روش چاکرابرتی است. آن­ها تعداد انجمن­های بین دو گام زمانی را متغیر درنظر گرفتند. همچنین فیلونو و پیزوتی در [[7](#_ENREF_7)] این کار را توسعه داده اند. آن­ها یک پارامتر ورودی به تابع هزینه اضافه کرده­اند که یک درجه کنترل برای متعادل نگه­داشتن همزمان دو پارامتر فراهم می­کند.

علیرضا قصبه و همکاران در [20] از ترکیب ایده­ی خوشه بندی کلونی مورچه با بهینه‌سازی کندوی زنبور عسل برای کشف انجمن در شبکه­های اجتماعی استفاده کرده­اند. خوشه‌بندی کلونی مورچه که یک جستجوی محلی است، توسط بهینه‌سازی کندوی زنبور عسل که یک راهکار سراسری است هدایت می‌شود. همچنین در پایان نامه آن­ها مدلی برای تخصیص زنبورهای رقصنده پیشنهاد شده است. راهکارهای پیشنهادی در کنار هم، باعث شده اند تا انجمن‌ها به طور دقیق تر و سریع تری کشف شوند.

# مراجع:

[1] J. Srivastava, M. A. Ahmad, N. Pathak, and D. K.-W. Hsu, "Data mining based social network analysis from online behavior," in *Tutorial at the 8th SIAM International Conference on Data Mining (SDM’08)*, 2008.

[2] S. Wasserman, *Social network analysis: Methods and applications* vol. 8: Cambridge university press, 1994.

[3] Y. Shang, P. Zhang, and Y. Cao, "A New Interest-sensitive and Network-sensitive Method for User Recommendation," in *Networking, Architecture and Storage (NAS), 2013 IEEE Eighth International Conference on*, 2013, pp. 242-246.

[4] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old: recommending people on social networking sites," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 201-210.

[5] J. DiMicco, D. R. Millen, W. Geyer, C. Dugan, B. Brownholtz, and M. Muller, "Motivations for social networking at work," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 2008, pp. 711-720.

[6] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *science,* vol. 328, pp. 876-878, 2010.

[7] F. Folino and C. Pizzuti, "An evolutionary multiobjective approach for community discovery in dynamic networks," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 26, pp. 1838-1852, 2014.

[8] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature,* vol. 446, pp. 664-667, 2007.

[9] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Statistical properties of community structure in large social and information networks," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 695-704.

[10] X. Dong-Fang and T. Chang-Shen, "Probabilistic Model Based Large-Scale Social Network Community Discovery Algorithm," in *Intelligent Computation Technology and Automation (ICICTA), 2014 7th International Conference on*, 2014, pp. 432-435.

[11] M. Coscia, F. Giannotti, and D. Pedreschi, "A classification for community discovery methods in complex networks," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* vol. 4, pp. 512-546, 2011.

[12] P. Pudil, J. Novovicova, and J. Kittler, "Simultaneous Learning of Decision Rules and Important Attributes for Classification Problems in Image-Analysis," *Image and Vision Computing,* vol. 12, pp. 193-198, Apr 1994.

[13] C. Pateritsas and A. Stafylopatis, "Memory-based classification with dynamic feature selection using self-organizing maps for pattern evaluation," *International Journal on Artificial Intelligence Tools,* vol. 16, pp. 875-899, Oct 2007.

[14] D. Windridge and J. Kittler, "Combined classifier optimisation via feature selection," *Advances in Pattern Recognition,* vol. 1876, pp. 687-695, 2000.

[15] J. A. Olvera-Lopez, J. A. Carrasco-Ochoa, J. F. Martinez-Trinidad, and J. Kittler, "A review of instance selection methods," *Artificial Intelligence Review,* vol. 34, pp. 133-143, Aug 2010.

[16] C. J. Whitaker, L. I. Kuncheva, and P. D. Cockcroft, "A logodds criterion for selection of diagnostic tests," *Structural, Syntactic, and Statistical Pattern Recognition, Proceedings,* vol. 3138, pp. 574-582, 2004.

[17] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis,* vol. 1, pp. 131-156, 1997.

[18] L. I. Kuncheva, "Fuzzy Rough Sets - Application to Feature-Selection," *Fuzzy Sets and Systems,* vol. 51, pp. 147-153, Oct 26 1992.

[19] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *Ieee Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, pp. 900-912, Jul 2004.

[20] T. Mitchell, *Machine Learning*: McGraw Hill, 1997.

[21] D. Ganley and C. Lampe, "The ties that bind: Social network principles in online communities," *Decision Support Systems,* vol. 47, pp. 266-274, 2009.

[22] J. Leskovec, K. J. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 631-640.

[23] M. E. Newman, "Detecting community structure in networks," *The European Physical Journal B-Condensed Matter and Complex Systems,* vol. 38, pp. 321-330, 2004.

[24] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2005, p. P09008, 2005.

[25] S. A. Moosavi and M. Jalali, "Community detection in online social networks using actions of users," in *Intelligent Systems (ICIS), 2014 Iranian Conference on*, 2014, pp. 1-7.

[26] P. Bródka, S. Saganowski, and P. Kazienko, "Group evolution discovery in social networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, 2011, pp. 247-253.

[27] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Link mining: models, algorithms, and applications*, ed: Springer, 2010, pp. 337-357.

[28] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu, "Graphscope: parameter-free mining of large time-evolving graphs," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 687-696.

[29] S. Asur, S. Parthasarathy, and D. Ucar, "An event-based framework for characterizing the evolutionary behavior of interaction graphs," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 3, p. 16, 2009.

[30] L. Tang, H. Liu, and J. Zhang, "Identifying evolving groups in dynamic multimode networks," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 24, pp. 72-85, 2012.

[31] T. Berger-Wolf, C. Tantipathananandh, and D. Kempe, "Dynamic community identification," in *Link Mining: Models, Algorithms, and Applications*, ed: Springer, 2010, pp. 307-336.

[32] D. Chakrabarti, R. Kumar, and A. Tomkins, "Evolutionary clustering," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 554-560.

[33] M.-S. Kim and J. Han, "A particle-and-density based evolutionary clustering method for dynamic networks," *Proceedings of the VLDB Endowment,* vol. 2, pp. 622-633, 2009.

1. Online [↑](#footnote-ref-1)
2. facebook [↑](#footnote-ref-2)
3. MySpace [↑](#footnote-ref-3)
4. Vertex [↑](#footnote-ref-4)
5. Node [↑](#footnote-ref-5)
6. Edge [↑](#footnote-ref-6)
7. Link [↑](#footnote-ref-7)
8. Link prediction [↑](#footnote-ref-8)
9. Temporal Smoothness [↑](#footnote-ref-9)
10. Snapshot Quality [↑](#footnote-ref-10)
11. Temporal Cost [↑](#footnote-ref-11)
12. Machine learning [↑](#footnote-ref-12)
13. DistanceMeasures [↑](#footnote-ref-13)
14. Information Measures [↑](#footnote-ref-14)
15. Dependence Measures [↑](#footnote-ref-15)
16. Correlation [↑](#footnote-ref-16)
17. Coefficient [↑](#footnote-ref-17)
18. Consistency Measures [↑](#footnote-ref-18)
19. Classifier Error Rate Measures [↑](#footnote-ref-19)
20. Generality [↑](#footnote-ref-20)
21. Machine Learning [↑](#footnote-ref-21)
22. Unsupervised Learning [↑](#footnote-ref-22)
23. Supervised Learning [↑](#footnote-ref-23)
24. Clustering [↑](#footnote-ref-24)
25. Unlabeled Data [↑](#footnote-ref-25)
26. Clustering [↑](#footnote-ref-26)
27. Regression [↑](#footnote-ref-27)
28. Classification [↑](#footnote-ref-28)
29. Classifier [↑](#footnote-ref-29)
30. Discriminative [↑](#footnote-ref-30)
31. Metaheuristic [↑](#footnote-ref-31)
32. Informed search [↑](#footnote-ref-32)