

**دانشگاه صنعتی شیراز**

**دانشکده مهندسی کامپیوتر و فناوری اطلاعات گروه نرم‌افزار**

**گزارش سمینار کارشناسی‌ارشد**

**در رشته مهندسی کامپیوتر گرایش شبکه**

بررسی روش های استخراج معنی از داده های شبکه های اینترنت اشیا با استفاده از روش های یادگیری ماشین روی داده های بزرگ

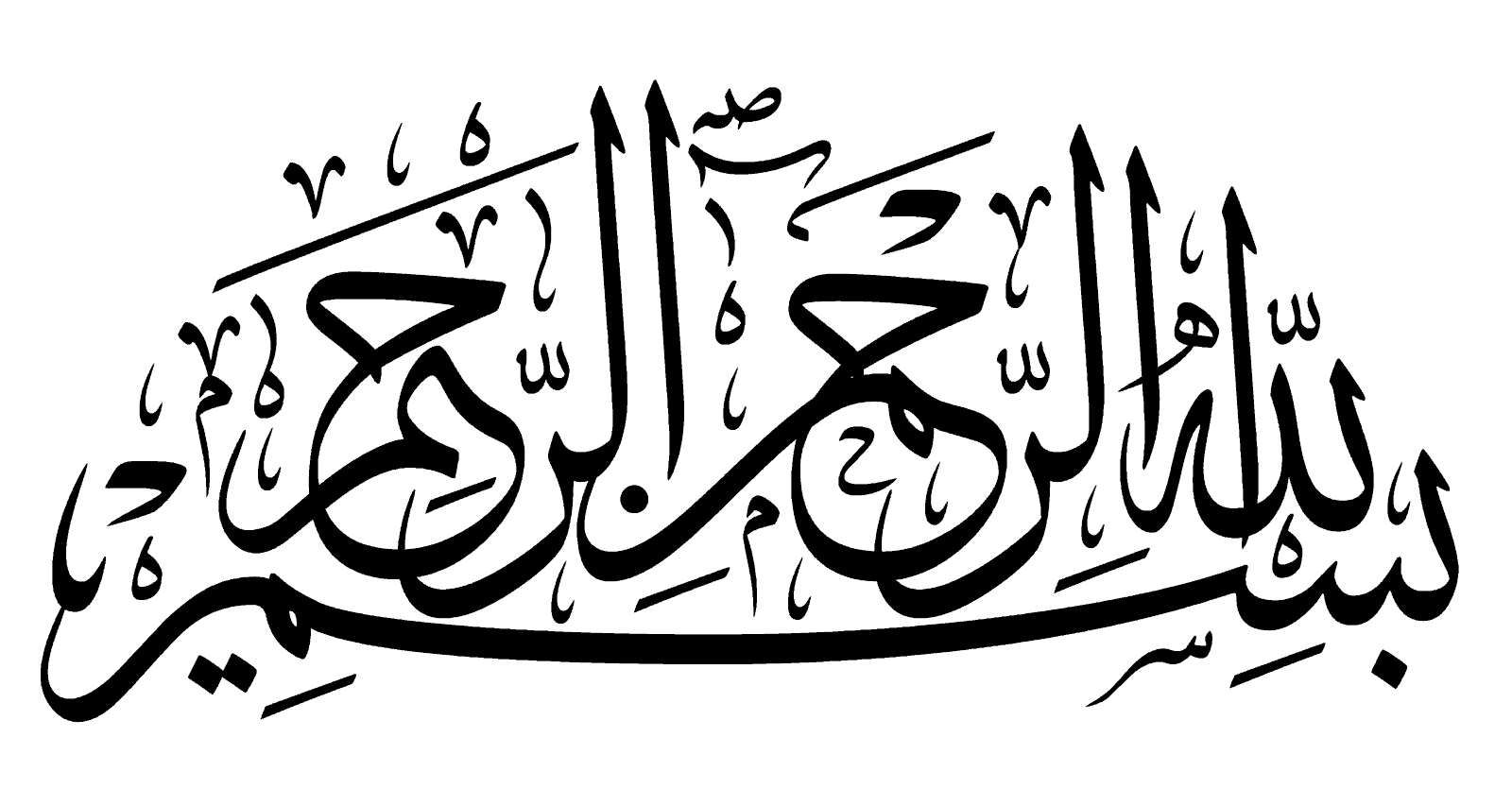
استاد راهنما:

دکتر جاویدان

استاد مشاور:

دکتر سرور

بهمن 97



چکیده

عنوان پایان نامه

نگارش :

یادگیری ماشین به معنی شناسایی الگوهای پیچیده و اتخاذ تصمیمات هوشمند به صورت خودکار می باشد. موضوعات مربوط به یادگیری ماشین را می توان در چهار گروه دسته بندی کرد. این موضوعات شامل یادگیری بی ناظر، یادگیری نیمه ناظر، یادگیری با ناظر و یادگیری فعال می باشد.

از داده کاوی به عنوان یک ابزار قدرتمند در استخراج دانش از داده ها استفاده می شود. با توجه اینکه حجم وسیعی از اطلاعات در حال تولید شدن می باشد به داده کاوی برای تحلیل داده ها و کشف دانش موجود در آن ها نیاز خواهیم داشت. داده کاوی در زمینه های مختلفی از جمله پزشکی، مهندسی و... کاربرد دارد. داده کاوی پزشکی در سال های اخیر مورد توجه زیادی قرار گرفته است. در ابتدا مفاهیم مربوط به یادگیری ماشین، داده کاوی و کاربرد آن در حوزه پزشکی را بررسی می کنیم و سپس برخی تحقیقات انجام شده در داده کاوی پزشکی در سال های اخیر را مورد بررسی قرار می دهیم.

**کلمات کلیدی :** هوش مصنوعی، یادگیری ماشین، داده کاوی، داده کاوی پزشکی

فهرست مطالب

2.[فصل دوم : مروری بر ادبیات تحقیق](#_Toc409474372) 9

[1-1- مقدمه](#_Toc409474373) 10

[1-2- تعاریف و ادبیات پایه مربوط به داده کاوی](#_Toc409474374) 10

[1-3- یادگیری ماشین](#_Toc409474377) 10

[1-4- یادگیری با ناظر](#_Toc409474378) 11

[1-4-1- دسته بندی 12](#ا32)

[1-4-1-1- الگوریتم های دسته بندی 13](#ک42)

[1-4-2- الگوریتم های مبتنی بر منطق 15](#د42)

[1-4-2-1- درخت تصميم 15](#د42)

[1-4-2-2- مجموعه آموزشی قوانین 17](#د42)

[1-4-3- تکنیک های مبتی بر پرسپترون](#_Toc409474383) 18

[1-4-3-1- شبکه هاي عصبي](#_Toc409474383) 18

[1-4-4- الگوریتم های آموزش آماری](#_Toc409474385) 21

[1-4-4-1- شبکه هاي بيزي](#_Toc409474385) 21

[1-4-5- آموزش مبتنی بر نمونه 22](#چ23)

[1-4-5-1- k نزدیک ترین همسایه 41](#چ23)

[1-4-5-2- ماشین بردار پشتیبان 23](#چ23)

[1-4-6- مقایسه روش ها](#_Toc409474387) 25

[1-5- یادگیری بدون نظارت](#_Toc409474389) 26

[1-5-1- خوشه بندی](#ت24) 27

[1-5-2- دسته بندی الگوریتم های خوشه بندی](#ت24) 27

[1-5-2-1- روش های افراز](#ت24) 28

[1-5-2-2- سلسله مراتبی](#ت24) 28

[1-5-2-3- مبتنی بر چگالی](#ت24) 28

[1-5-2-4- مبتنی بر گرید](#ت24) 29

[1-5-2-5- مبتنی بر مدل](#ت24) 29

[1-5-3- الگوریتم های خوشه بندی](#_Toc409474397) 31

[1-5-3-1- الگوریتمFuzzy C-Means(FCM)](#_Toc409474397) 31

[1-5-3-2-الگوریتم BIRCH](#_Toc409474397) 31

[1-5-3-3-الگوریتم DENCLUE](#_Toc409474397) 32

[1-5-3-4-الگوریتم Optimal Grid (OptiGrid)](#_Toc409474397) 33

[1-5-3-5-الگوریتمEXPECTATION-MAXIMIZATION (EM)](#_Toc409474397) 33

[1-5-4- بررسی معیارهای سنجش روش های خوشه بندی](#ت24) 34

[1-5-4-1- نوع ديتاست](#_Toc409474397) 34

[1-5-4-2- سايز ديتاست](#_Toc409474397) 34

[1-5-4-3- پارامترهاي ورودي](#_Toc409474397) 35

[1-5-4-4- مديريت داده هاي پرت و نويز](#_Toc409474397) 35

[1-5-4-5- پيچيدگي زماني](#_Toc409474397) 35

[1-5-4-6- پایداری](#_Toc409474397) 35

[1-5-4-7- چگونگی برخورد با داده های حجم بالا](#_Toc409474397) 35

[1-5-4-8- شکل خوشه](#_Toc409474397) 36

[1-6- نتیجه گیری](#_Toc409474397) 36

3.فصل [سوم : مروری بر تحقیقات اخیر انجام شده](#_Toc409474398) 38

[2-1- مقدمه](#_Toc409474399) 39

[2-2- تحقیقات جدید انجام شده در داده کاوی پزشکی 39](#ن25)

[مراجع](#_Toc409474402) 49

فهرست شکل‌ها

[شکل 1-1 دسته بندی روش های خوشه بندی 29](#ش4)

فهرست کلمات اختصاری

ANFIS Adaptive Neuro Fuzzy Inference System

SEER the Surveillance, Epidemiology, and End Results

SVM Support Vector Machine

aiNet-K Artificial Immune Network + k-means

ANOVA ANalysis Of VAriance

UCI University of California

M-CLUBS Microarray data CLustering Using Binary Splitting

KEEL Knowledge Extraction based on Evolutionary Learning

MTD-KNN Mega Trend Diffusion – K-Nearest Neighbor

MTD-SVM Mega Trend Diffusion – Support Vector Machine

CPAR Classification Based on Predictive Association Rule

BIRCH Balanced Iterative Reducing and Clustering using Hierarchies

DENCLUE Clustering Based on Density Distribution Functions "DENsity-based CLUstEring

C&RT Classification and Regression Trees

CHAID CHi-squared Automatic Interaction Detection

NCBI National Center for Biotechnical Information

GEO Gene Expression Omnibus

DNF Disjunctive Normal Form

EM Expectation Maximization

ROC Receiver operating characteristic

Opt-aiNET An Immune Network Algorithm for Optimisation

PSO Particle swarm optimization

OptiGrid Optimal Grid

FCM Fuzzy C-Means

**فصل دوم: مروري بر ادبیات تحقیق**

* 1. مقدمه

در این فصل ابتدا به تعریف یادگیری ماشین و داده کاوی پرداخته می­شود. سپس به صورت خاص‌تر، تعریف داده کاوی بیان می‌گردد.

* 1. تعاریف و ادبیات پایه مربوط به داده کاوی

داده کاوی[[1]](#footnote-1)چیست؟

در لغت به معنی بررسی و کاوش داده ها می باشد.

از لحاظ مفهومی داده کاوی عبارت است از:

* ترکیب روش های تحلیل داده ها و الگوریتم های پیشرفته برای پردازش داده های با حجم بالا می باشد. داده کاوی امکان تحلیل نوع داده های جدید و همچنین روش های جدیدی برای تحلیل نوع داده های قدیمی به ما می دهد ]1[.
* فرآیندی است که به کمک کامپیوتر برای کاوش و تحلیل مجموعه های عظیم داده ها انجام می شود تا مفهموم موجود در داده ها را استخراج کند [2].
  1. یادگیری ماشین

یادگیری ماشین بستري براي دانشمندان براي کاوش مدل هاي و الگوريتم هاي براي يادگيري ماشين از داده ها فراهم کرده است [3]. ميزان داده هاي توليد شده به دليل رشد تکنولوژي هايي مانند پردازش هاي ابري و سنسور ها به شدت افزايش يافته است. به دليل اين گسترش سايز در داده ها و پيچيدگي هاي پردازشي که اين داده ها با خود دارند دانشمندان و مهندسين نياز حياتي تري به اين فيلد تحقيقاتي براي خودکار سازي فرآيندهاي تصميم گيري پيچيده و حل مسائل مربوط به آناليز اين داده ها دارند.

دليل توجه به يادگيري ماشين استخراج اطلاعات از داده ها و استفاده از کامپيوترها براي اتخاذ تصميمات پيچيده مي باشد به عنوان مثال قوانين تصميم گيري توسط يادگيري ماشين با سرعت مناسب از داده ها به دست آمده و اين قوانين نيز توسط سيستم هاي کامپيوتري براي تصميم گيري استفاده مي شوند. به عنوان مثال مي توان به تکنولوژي DNA microarray اشاره کرد که به بايولوژيست ها و متخصصين پزشکي اجازه بررسي هزاران ژن يک نمومه بافت را مي دهد و به آن ها امکان شناسايي ژن هاي سرطاني در يک مطالعه سرطان مي دهد. گرچه اطلاعات توليد شده از يک آزمايش DNA microarray و يا دستگاه هاي اندازه گيري ديگر را به دليل سايز و پيچيدگي زيادي که دارد نمي توان به صورت دستي پردازش يا آناليز کرد. در مطالعات مرتبط با سرطان يادگيري ماشين به عنوان يک ابزار ارزشمند براي تشخيص ژن هاي سرطاني از ميان هزاران ژن مي باشد[4].

تکنيک هاي یادگیری ماشين بر اساس مشکلاتي که مي توانند حل کنند به دو دسته، شامل يادگيري با نظارت[[2]](#footnote-2) و يادگيري بدون[[3]](#footnote-3) نظارت تقسيم مي شوند. در ادامه این بخش به بررسي اين تکنيک ها خواهيم پرداخت.

* 1. يادگيري با ناظر

الگوريتم هاي يادگيري نظارت شده سعي در يادگيري رابطه ورودي-خروجي (رابطه تابعي يا وابستگي) با استفاده از يک ديتاست آموزشي دارد. ديتاست ورودي شامل n ورودي که هر کدام از ورودي ها به صورت زوج هايي از x و y مي باشد است که ورودي x خود شامل يک بردار m بعدي است و y بيانگر برچسب داده x (پاسخ سيستم) مي باشد.y براي مسائل دسته بندی [[4]](#footnote-4) گسسته و براي مسائل رگراسيون به صورت پيوسته مي باشد.

ماشين هاي بردار پشتيبان [[5]](#footnote-5) و شبکه هاي عصبي مصنوعي را مي توان به عنوان نمونه هايي از اين روش نام برد.

به طور کلي دو نوع الگوريتم يادگيري با نظارت وجود دارد :

1. دسته بندی
2. رگرسیون

اين دو روش را در ادامه توضيح داده شده است.

* + 1. دسته بندی

در مسائل دسته بندي مجموعه آموزش شامل نمونه هايي از کلاس هاي مختلف مي باشد. ساده ترين مسئله دسته بندي، دسته بندي دودويي مي باشد که مجموعه آموزش تنها شامل داده هايي از دو کلاس متفاوت مي باشد. خروجي الگوريتم مي باشد که بيانگر کلاسي که بردار ورودي که در دسته بندي ما به آن تعلق دارد مي باشد.

بردار ورودي شامل ويژگي ها يا صفاتي است که براي دسته بندي ورودي هاي X مختلف استفاده مي شود است. پروسه يادگيري در مسائل دسته بندي شامل ايجاد طبقه گر هايي است که با استفاده از آن ها بتوان نمونه هايی از X که تا به حال نديده ايم، را براي ما دسته بندي کند. به عبارت ديگر ماشين در ابتدا از داده هاي آموزش تربيت شده و در مراحل بعدي براساس يادگيري که داشته تصميم گيري می کند. در مسائل دسته بندي چند کلاسه چندين طبقه گر باينري ايجاد شده و براي پيش بيني برچسب داده هاي جديد استفاده مي شوند.

به عنوان مثال يک مسئله N کلاسه درحالت کلي به N مسئله دسته بندي دودويي تبديل مي شوند. مسائل دسته بندي را مي توان در حوزه هاي مختلفي مانند تشخيص اشياء، تشخيص دست خط، دسته بندي متن، آناليز بيماري و مطالعات مرتبط به DNA microarray نام برد. واژه نظارت شده از اين مسئله گرفته شده است که داده هاي مجموعه آموزش داراي برچسب هايي که مشخص کننده کلاس آن داده می باشد. اين داده ها به عنوان معلمي براي تربيت الگوريتم يادگيري مي باشد.

در مسئله رگراسيون، وظيفه الگوریتم یادگیری پيدا کردن نگاشتي بين ورودي X و خروجي Y مي باشد. خروجي Y در مسئله رگراسيون برخلاف مسئله دسته بندي که گسسته است پيوسته مي باشد.

فرآيند يادگيري در رگراسيون شامل پيدا کردن تابعي با تعداد ورودي به تعداد ويژگي هاي موجود در وکتور X و يک خروجي اسکالر Y مي باشد.

مسائل رگراسيون را مي توان در حوزه هاي مختلفي مانند پيش بيني سري هاي زماني، سيستم هاي کنترل، سيستم هاي جهت يابي و آناليز نرخ سود در سيستم هاي مالي استفاده کرد

* + - 1. الگوريتم هاي دسته بندی

يادگيري ماشين استنتاجي[[6]](#footnote-6) پروسه يادگيري مجموعه اي از قوانين، از نمونه هاي داده های موجود در مجموعه آموزش مي باشد. به طور کلي مي توان گفت ايجاد يک طبقه گر که بتواند داده هاي ورودي جديد را براي ما به درستي دسته بندي کند.

مرحله اول در اين فرآيند جمع آوري داده ها مي باشد. اگر يک متخصص در زمينه مورد نظر براي مشخص کردن ويژگي هاي مهم در داده ها موجود بود، مي توان از آن کمک گرفت، در غير اين صورت مي توان از روش بروت فورس [[7]](#footnote-7) به عنوان ساده ترين راه استفاده کرد. به اين معني که کليه ويژگي ها را به اين اميد که ويژگي ها ايزوله باشند اندازه گيري کرد. البته ديتاست هايي که به اين روش جمع آوري مي شوند به دليل دارا بودن داده هاي نويز و مقادير ويژگي هاي خالي مناسب استفاده به صورت مستقيم نمي باشند. به همين دليل نيازمنده انجام پيش پردازش داده هاي ديتاست مي باشيم[5].

سلسله مراتبي از مشکلات وجود دارند که در آماده سازي داده ها[[8]](#footnote-8) و پيش پردازش[[9]](#footnote-9) آن ها مي توانند به وجود بيايند :

- وجود مقادير غيرممکن در داده هاي ورودي

- براي برخي ويژگي ها[[10]](#footnote-10) در برخي از نمونه ها داده اي وارد نشده است.

- وجود ويژگي هاي نامرتبط درون نمونه هاي داده

وجود داده هاي نويز و داده های پرت بايد توسط نرم افزار هاي مديريت داده ها بررسي شوند. به صورت ايده آل در مواردي که داده هايي موجود نباشد، مي توان آن ها را مجددا وارد کرد.

در صورتي که امکان ورود مجدد داده ها وجود نداشته باشد مسئله پر کردن مقادير در دسته مسائل داده هاي گم شده قرار مي گيرد که ساده ترين راه حل براي اين مسئله حذف داده ها مي باشد.

هاج و آستین در [6] مروري بر تکنيک هاي رفع داده هاي نويز داشته اند.

داده هاي ناقص يک مشکل اجتناب ناپذير در مقابله با بيشتر منابع داده جهان واقعي مي باشد. عموما چند فاکتور مهم در هنگام پردازش داده هاي ويژگي هاي نامشخص در نظر گرفته مي شود. يکي از مهمترين علت وجود اين داده ها مشکل عدم شناخت مي باشد :

- داده ها به اين دليل که از بين رفته يا گم شده اند وجود ندارند.

- براي برخي از نمونه های داده، وجود برخی از ويژگي ها ممکن نيست.

- براي يک نمونه داده، طراح ديتاست آموزش اهميتي به مقادیر برخي ويژگي ها نداده و به اصطلاح آن ويژگي ها را ويژگي هاي بي اهميت[[11]](#footnote-11) در نظر گرفته است.

با توجه به شرايط مختلف محققان روش هاي مختلفي براي مديريت داده های گم شده انتخاب مي کنند [7].

انتخاب زيرمجموعه اي از ويژگي ها، پروسه مشخص کردن و حذف ويژگي هاي اضافه و نامرتبط با داده ها مي باشد[8]. اين مسئله براي کاهش ابعاد داده ها و افزايش سرعت اجرا و عملکرد بهتر الگوريتم ها استفاده مي شود. اين موضوع که تعداد زيادي از ويژگي ها به ديگر ويژگي ها وابستگي دارند اغلب کارايي الگوريتم هاي يادگيري ماشين را تحت تاثير خود قرار مي دهند. اين مسئله به عنوان ايجاد مجموعه ويژگي هاي جديد، از ويژگي هاي موجود داده ها مي باشد [9].

انتخاب الگوريتم يادگيري مناسب براي کاربرد موردنظر يکي از حياتي ترين مراحل فرآیند یادگیری ماشین مي باشد. ارزيابي دسته بندی کننده [[12]](#footnote-12) معمولا توسط اندازه گيري ميزان دقت انجام مي گيرد، دقت يک دسته بندی کننده برابر با تعداد پيش بيني درست انجام داده تقسيم بر تعداد کل پيش بيني ها است.

سه تکنيک براي سنجش دقت يک دسته بندی کننده وجود دارد. تکنيک اول تقسيم بندي مجموعه آموزش به اينصورت که دو سوم داده به آموزش و يک سوم به ارزيابي اختصاص داشته باشد.تکنيک ديگر به روش Cross-Validation معروف است، در اين روش مجموعه آموزش به n قسمت افراز شده و به ازاي هر بخش از افراز دسته بندی کننده روي باقي بخش ها آموزش مي بيند، ميانگين خطاي هر زير مجموعه معياري براي سنجش خطاي دسته بندی کننده است. روش سوم Leave-One-out مي باشد، این روش يک نمونه خاص از روش Cross-Validation است به اين صورت که کليه زيرمجموعه هاي مجموعه تست شامل تنها يک عضو مي باشند، اين نوع اعتبارسنجي از نظر هزينه محاسباتي دارای بيشترين هزينه است، این روش در مواقعي که تخمين دقيقتري از ميزان خطاي دسته بندی کننده مهم مي باشد، روشي مناسب می باشد.

دسته بندي نظارت شده يکي از وظايف متداول سيستم هاي هوشمند مي باشد. بنابراين تکنيک هاي مختلفي براي دسته بندي بر اساس هوش مصنوعي مانند تکنيک هاي بر پايه منطق و تکنيک هاي برپايه پرسپترون [[13]](#footnote-13) همچنين تکنيک هاي مبتني بر آمار مانند شبکه هاي بيزي [[14]](#footnote-14) و تکنيک هاي مبتني بر نمونه [[15]](#footnote-15) ايجاد شده اند.

در ادامه اين بخش به بررسي مهمترين روش هاي يادگيري با نظارت خواهيم پرداخت.

* + 1. الگوریتم های مبتنی بر منطق [[16]](#footnote-16)

در اين بخش به بررسي روش هاي مبتني بر منطق درخت تصميم و تکنيک هاي مبتني بر قانون مي پردازيم.

* + - 1. **درخت تصميم**

مورتي در [10] به مروري بر روش هاي درخت تصميم و مثال هايي از کاربرد آن در يادگيري ماشين پرداخته است.

درخت هاي تصميم، درخت هايي هستند که نمونه ها را براساس مرتب سازي مقادير ويژگي هايشان دسته بندي مي کنند. نمونه ها با شروع از ريشه و بر اساس مقادير ويژگي دسته بندي مي شوند. هر گره در درخت تصميم بيانگر يک ويژگي که براي دسته بندي استفاده شده است مي باشد. هر شاخه در درخت بيانگر مقدار براي هر ويژگي مي باشد. نمونه ها با شروع از ريشه دسته بندي مي شوند و بر اساس مقادير مربوط به ويژگي هايشان مرتب مي شوند.

مسئله ايجاد درخت تصميم بهينه يک مسئله NP سخت [[17]](#footnote-17) مي باشد و به همين دليل تئوريسن ها به دنبال روش هايي مناسب براي ايجاد اين درخت بوده اند.

ويژگي که به بهترين نحو داده هاي آموزش را تقسيم مي کند به عنوان ريشه قرار مي گيرد.در [10] روش هاي متفاوتي براي پيدا کردن بهترين ويژگي براي انتخاب نود ريشه معرفي شده اند.

در صورتي که دو درخت روي داده هاي آموزش دقت پيش بيني يکساني ارائه دهند درختي که داراي تعداد گره هاي کمتر است کانديد مناسب تري براي انتخاب مي باشد. بيشتر الگوريتم ها از روش هايي براي هرس درخت استفاده مي کنند. در [11] روش هاي هرس مختلف با هم مقايسه شده اند.

بهترين الگوريتم براي ايجاد درخت تصميم در تحقيقات ارائه شده الگوريتم C4.5 مي باشد [12].

با مطرح شدن داده های حجم بالا چالش ها روز به روز افزايش يافته و روش هاي قديمي ايجاد درخت تصميم داراي محدوديت هايي مي باشند. اولين محدوديت ايجاد درخت تصميم مي باشد. ايجاد درخت تصميم در مواقعي که ديتاست بزرگ مي باشد پروسه اي زمانبر مي باشد. همچنين اگرچه با استفاده از موازي سازي مي توان بهبود هايي در ايجاد درخت توسط الگوريتم ها ايجاد کرد اما مسئله توزيع داده ها بر روي سيستم ها براي اجراي موازي يک مسئله مهم مي باشد.

براي حل اين چالش ها الگوريتم با استفاده از مدل محاسباتي Map Reduce پياده سازي شده است و در زماني که ديتاست بسيار بزرگ مي باشد اين الگوريتم را مي توان استفاده نمود [13].

* + - 1. **مجموعه آموزشی قوانین**

درخت تصميم را مي توان به صورت مجموعه اي از قوانين ترجمه کرد به اين صورت که براي هر شاخه درخت، از ريشه تا يکي از برگ ها قانوني مشخص کنیم.

يک روش ديگر براي استخراج قوانين استفاده از الگوريتم هايي مي باشد که بدون تشکيل درخت تصميم قوانين را از داده ها استخراج مي کنند

هر کلاس دسته بندي توسط يک DNF [[18]](#footnote-18) بيان مي شود. يک K-DNF به صورت زير بيان مي شود :

در اين مثال k بيانگر تعداد قوانين مجزا و n مشخص کننده تعداد ترکيب عطفي در هر قانون مي باشد. هدف ايجاد کمترين مجموعه قوانين پوشش دهنده مجموعه آموزش است. در صورت زياد شدن تعداد قوانين مي توان اين نتيجه رسيد که درخت تصميم به جاي سعي در پياده سازي فرض ها و يادگيري قوانين سعي در حفظ کردن مجموعه آموزش دارد و به اصطلاح آموزش بیش از حد [[19]](#footnote-19) اتفاق افتاده است.

يک الگوريتم جداسازي و غلبه در پيدا کردن قوانين ابتدا به دنباله قانوني براي توصيف بخشي از داده هاي مجموعه آموزش مي گردد و اين داده ها را جدا کرده و به صورت بازگشتي براي بقيه داده ها به دنبال قانون مي گردد و اين عمل را تا پوشش دادن کل داده ها انجام مي دهد.

براي سيستم هاي توليد قانون، توليد قوانين تصميمي داراي پيش بيني بالا، مهم مي باشد. ان و سرکون مروري بر روش هاي آماري و مبتني برتجربه براي سنجش کيفيت قوانين داشته است [14].

* + 1. تکنیک های مبتنی بر پرسپترون [[20]](#footnote-20)

يکي ديگر از روش هاي معروف در دسته بندي استفاده از پرسپترون ها مي باشد.پرسپترون را مي توان به اين صورت توصيف کرد :

اگر تا ويژگي هاي ورودي باشند و تا وزن هاي بردارهاي پيش بيني باشند که خروجي پرسپترون به صورت مي باشد.

با بررسي يک حد آستانه اگر مجموع محاسبه شده بالاتر از آن باشد خروجي 1 و در غير اينصورت 0 مي باشد.

مرسوم ترين روش استفاده از الگوريتم پرسپترون براي يادگيري از داده هاي آموزش دسته اي مي باشد. اين الگوريتم چندين بار تکرار شده تا بردار پيش بيني کننده اي که براي کليه داده هاي مجموعه آموزش درست باشد بدست آورد.

* + - 1. شبکه هاي عصبي [[21]](#footnote-21)

پرسپترون تنها مي تواند داده هايي که به صورت خطي قابل جدا سازي باشند را جدا کند. اگر بتوان داده ها را توسط يک خط يا صفحه جدا کرد آن گاه داده هاي ورودي به صورت خطي جدايي پذير بوده و پرسپترون جواب را پيدا مي کند.

اگر داده ها به صورت خطي جدايي پذير نباشند پرسپترون هيچ گاه به نقطه اي که داده ها درست دسته بندي شده اند نخواهد رسيد. شبکه هاي عصبي مصنوعي براي حل اين مسئله ايجاد شده اند. ژانگ در [15] مروري بر مطالعات انجام شده در حوزه شبکه هاي عصبي مصنوعي و دسته بندی انجام داده است.

يک شبکه عصبي چند لايه از تعداد زيادي نورون که با يکديگر در ارتباط مي باشند ايجاد شده است. واحدهاي موجود در يک شبکه عصبي مصنوعي را مي توان در سه کلاس دسته بندي کرد : واحد ورودي، که داده ها براي پردازش به آن ها تحويل داده مي شوند. واحد خروجي، واحدهايي که نتايج پردازش در آن ها موجود مي شود و واحد هايي که بين واحد ورودي و خروجي قرار گرفته اند واحد هاي مياني يا پنهان مي باشند.

شبکه هاي عصبي پیش خور به سيگنال اجازه عبور در يک مسير يک طرفه را مي دهد اين مسير از ورودي شروع و به خروجي مي رسد. در ابتدا شبکه روي زوج داده هاي ورودي-خروجي آموزش دیده سپس از آموزش به عنوان دسته بندی کننده براي داده هاي جديد استفاده مي شود.

عموما تشخيص سايز لايه مخفي مسئله مي باشد، تخمين تعداد لايه هاي کم مي تواند باعث تخمين ضعيف نتایج شده و زياد کردن لايه هاي پنهان مي توان باعث مشکل Overfitting شود. در [16] مي توان يک استدلال خوب در اين موضوع را پيدا کرد.

کم و پلاسکتا نيز مينيمم تعداد نورون ها و تعداد نمونه هاي لازم براي برنامه ريزي يک کار در شبکه هاي عصبي پیش خور را بررسي کرده اند [17] .

شبکه هاي عصبي مصنوعي وابسته به سه جنبه اساسي مي باشند. ورودي و توابع فعال ساز، معماري شبکه عصبي و وزن هر ورودي. با توجه به اينکه دو مورد اول ثابت مي باشد رفتار يک شبکه عصبي توسط مقادير جاري وزن ها تعريف مي شود. وزن هاي شبکه عصبي قبل از آموزش معمولا به مقادير تصادفي تنظيم مي شوند. ورودي ها به شبکه داده مي شود و خروجي هاي حاصل با خروجي هاي مورد نظر مقايسه مي شوند و کليه وزن ها در شبکه در جهتي که شبکه را به سمت ايجاد خروجي مناسب هدايت کند تغيير مي دهند. الگوريتم هاي متفاوتي براي آموزش يک شبکه عصبي وجود دارد [18]. معروف ترين الگوريتم آموزش شبکه هاي عصبي براي تخمين وزن هاي شبکه الگوريتم پس انتشار [[22]](#footnote-22) مي باشد.

شبکه هاي عصبي پیش خور [[23]](#footnote-23) معمولا با استفاده از الگوريتم پس انتشار آموزش داده مي شوند. مشکلي که در آموزش وجود دارد اين الگوريتم براي اکثر کاربردها بسيار کند مي باشند. يکي از راه هاي سرعت بخشيدن به فرآيند آموزش تخمين وزن هاي اوليه به جاي استفاده از وزن هاي تصادفي مي باشد [19].

مي توان از الگوريتم ژنتيک يا ديگر الگوريتم هاي فرامکاشفه اي براي آموزش شبکه هاي عصبي و پيدا کردن ساختار مناسب شبکه استفاده کرد [20].

اگرچه شبکه هاي عصبي چند لايه و درخت هاي تصميم دو روش کاملا متفاوت در دسته بندی مي باشند برخي از محققان مقايسه اي ما بين اين روش ها انجام داده اند [21]. نتايج اين مطالعه به اين شرح است :

- شبکه هاي عصبي معمولا در فراهم سازي يادگيري افزايشي نسبت به درخت هاي تصميم موفق تر مي باشند.

- زمان آموزش يک شبکه عصبي معمولا طولاني تر از زمان آموزش درخت تصميم مي باشند.

- شبکه هاي عصبي معمولا به خوبي درخت تصميم کار مي کنند و به ندرت بهتر از درخت تصميم جواب مي دهند.

* + 1. الگوریتم های آموزش آماری

برخلاف شبکه هاي عصبي مصنوعي روش هاي مبتني بر علم آمار داراي يک مدل آماري مشخص که بيانگر احتمال تعلق يک نمونه داده به يک کلاس مي باشد. در اين دسته از الگوريتم هاي دسته بندی مي توان به شبکه هاي بيزي و متدهاي يادگيري مبتني بر نمونه اشاره کرد.

يک کتاب مناسب براي شبکه هاي بيزي توسط Jensen [22] ارائه شده است. بنابراين در اين بخش به مرور برخي از مطالعات در اين زمينه خواهيم پرداخت.

* + - 1. **شبکه هاي بيزي**

يک شبکه بيزي يک مدل گرافيکي براي نشان دادن رابطه احتمال ميان مجموعه اي از ويژگي ها مي باشد. ساختار شبکه بيزي S يک گراف جهت دار بدون دور مي باشد و نودهاي S نظير به نظير با ويژگي X در ارتباط مي باشند. يال ها بيانگر رابطه وابستگي ميان ويژگي ها مي باشد همچنين عدم وجود يال بيانگر استقلال شرطي مي باشد.

جذاب ترين ويژگي شبکه هاي بيزي در مقايسه با درخت هاي تصميم و شبکه هاي عصبي امکان در نظر گرفتن اطلاعات قبلي در مورد يک مساله است همچنین مشکلي که شبکه هاي بيزي دارند مناسب نبودن براي داده هايي با تعداد ويژگي بالا مي باشد. به اين دليل که تلاش براي ايجاد يک شبکه خيلي بزرگ از نظر فضايي و زماني امکان پذير نيست.

شبکه هاي بيزي ساده شبکه هايي مي باشند که گراف هاي آن داراي يک نود ريشه و تعدادي فرزند است و اين فرض که بين نود هاي فرزند استقلال وجود دارد برقرار است. ويژگي اصلي شبکه هاي بيزي ساده سريع بودن فرآيند آموزش آن ها مي باشد. اگر يک ويژگي عددي باشد معمولا در کليه الگوريتم هاي بيزي آن ويژگي در پيش پردازش داده ها را گسسته مي کنند.

* + 1. آموزش مبتنی بر نمونه [[24]](#footnote-24)

يک نوع ديگر از روش هاي آماري يادگيري بر اساس نمونه ها مي باشد. الگوريتم هاي آموزش مبتنی بر نمونه الگوريتم هاي يادگيري تنبل مي باشند. اين الگوريتم ها پروسه استنتاج را تا زمان انجام دسته بندي به تاخير مي اندازند. الگوريتم هاي يادگيري تنبل به زمان محاسباتي کمتري در فاز آموزش نسبت به ديگر الگوريتم ها مانند درخت يادگيري و يا شبکه هاي عصبي نياز دارند همچنین به زمان بيشتري در زمان انجام دسته بندي نياز دارند.

يکي از سر راست ترين روش هاي يادگيري مبتني بر نمونه الگوريتم نزدیک ترین همسایه مي باشد.

* + - 1. **k نزدیک ترین همسایه [[25]](#footnote-25)**

الگوريتم k نزدیک ترین همسایه بر اساس اين اصل که داده هاي نزديک به هم در يک ديتاست داراي شباهت در خصوصيات مي باشند کار مي کند. اگر نمونه ها داراي برجسب باشند داده هاي بدون برچسب را مي توان در کلاس نزديک ترين داده به آن دسته بندي کرد. k نزدیک ترین همسایه با بررسي k عنصر داده نزديک به داده مورد سوال آن داده را در بيشترين تکرار کلاس در k نمونه دسته بندي مي کند.

به صورت عمومي، هر نمونه را مي توان يک نقطه در يک فضاي n بعدي در نظر گرفت که هر کدام از ابعاد توصيف کننده مقدار يکي از n ويژگي هاي نمونه مي باشد. مکان هر نقطه در فضا به اندازه فاصله نسبي آن از ديگر نقاط داراي اهميت نمي باشد. هدف کمينه سازي فاصله ميان نمونه هاي يک کلاس و بيشينه سازي فاصله ما بين نمونه هاي موجود در کلاس هاي مختلف است.

قدرت روش k نزدیک ترین همسایه در کاربردهاي واقعي بسياري اثبات شده است اما استثناهايي در رابطه با کارکرد مناسب k نزدیک ترین همسایه وجود دارد :

- روش k نزدیک ترین همسایه به فضاي ذخيره سازي زيادي نياز دارد.

- اين روش به تابع تشخيص شباهتي که استفاده شده حساس است.

- مشکل در تعيين مقدار مناسب براي k

- هزينه محاسباتي بالا

اوکاموتو و یوگامی دقت مورد انتظار k نزدیک ترین همسایه را به عنوان تابعي از ويژگي هاي دامين مورد استفاده مانند تعداد نمونه هاي موجود در مجموعه آموزش و تعداد خصوصيات مرتبط و نامرتبط، نرخ نويز، احتمال مربوط به هر خصوصيت و مقدار k نشان داده اند [23].

همانگونه که در قبل اشاره شد يکي از معايب روش هاي مبتني بر نمونه ميزان هزينه محاسباتي بالا در آن ها مي باشد. يک مسئله کليدي در بسياري از کاربردها تشخيص ويژگي هايي که در مدل سازي بايد استفاده شود توسط روش هاي انتخاب ويژگي تعيين مي شود [8]. انتخاب ويژگي مي تواند به بهبود دقت و مقياس پذيري و همچنين کاهش زمان دسته بندي کمک کند. مساله ديگر انتخاب نمونه هايي که براي مدل سازي بايد استفاده شوند مي باشد [24].

* + - 1. ماشین بردار پشتیبان [[26]](#footnote-26)

ماشين هاي بردار پشتيبان يا بردار پشتیبان يکي از تکنيک هاي جديدتر در زمينه يادگيري نظارت شده در يادگيري ماشين مي باشد.

هدف ماشين هاي بردار پشتيبان ايجاد يک مارجين فوق صفحه براي جداسازي دو کلاس داده از يکديگر است. بيشينه سازي مارجين و ايجاد بيشتر فاصله توسط فوق صفحه ما بين کلاس هاي داده تابع هدف اين روش می باشد. در واقع ماشين هاي بردار پشتيبان بر خلاف روش هاي ديگر مانند شبکه هاي عصبي مصنوعي که به دنبال کمينه سازي خطا مي باشند اين روش به دنبال کمينه سازي ريسک دسته بندی مي باشد.

در نمونه داده هايي که به صورت خطي جدايي پذير مي باشند هنگامي که فوق صفحه جداکننده بدست آمد داده هايي که در هر سمت مارجين اين فوق صفحه قرار دارند به عنوان نقاط پشتيبان شناخته مي شوند. راه حل خروجي اين روش شامل ترکيب خطي اين نقاط مي باشد.

پيچيدگي روش ماشین بردار پشتیبان تحت تاثير تعداد ويژگي هاي داده هاي آموزش قرار ندارد. به همين دليل ماشین بردار پشتیبان براي یادگیری با داده ها آموزش با تعداد ويژگي بالا مورد مطالعه قرار مي گيرند.

از آنجا که انتخاب ماکسيمم مارجين در ماشین بردار پشتیبان به آن اجازه انتخاب ميان چندين فوق صفحه کانديد مي دهد. براي بيشتر ديتاست ها به دليل وجود داده هايي که درست دسته بندي نشده اند. ممکن است ماشین بردار پشتیبان نتواند هيچ فوق صفحه اي را پيدا کند. اين مسئله مي تواند با استفاده از مارجين هاي نرم که تحمل پذيري در برابر دسته بندي اشتباه دارند برطرف شود [25].

بنابراين اکثر مسائل دنياي واقعي شامل داده هاي جدايي ناپذير مي باشند و نمي توان هيچ فوق صفحه اي که داده هاي مثبت را از منفي جداسازي کند يافت. يک راه حل براي اين مشکل نگاشت داده ها به يک فضا با ابعاد بيشتر است و تعريف يک فوق صفحه در آن فضا است. اين فضا با ابعاد بالاتر را فضاي ويژگي نام دارد.

با استفاده از يک فضاي ويژگي مناسب مي توان هرگونه داده اي را جدا نمود.

استفاده از توابع کرنل مي توان داده ها را بدون انتقال به فضاي ويژگي مستقيما در فضاي ويژگي محاسبه نمود. وقتي که يک فوق صفحه ايجاد شد از تابع کرنل براي نگاشت داده هاي جديد در فضاي ويژگي و دسته بندي استفاده مي شود.

انتخاب مناسب تابع کرنل به دليل تعيين فضاي ويژگي که داده ها در آن دسته بندي مي شوند بسيار مهم مي باشد. يک راه حل مرسوم در انتخاب کرنل تخمين بازه اي از تنظيمات ممکن و استفاده از Cross-Validation روي داده هاي آموزش براي انتخاب بهترين کرنل است. به دليل استفاده از اين روش يکي از محدوديت هاي ماشین بردار پشتیبان سرعت پايين در آموزش است. انتخاب کرنل مي تواند توسط روش هاي استفاده شده براي تعيين تعداد لايه هاي شبکه هاي عصبي نیز انجام شود.

آموزش ماشین بردار پشتیبان با حل مسئله برنامه ریزی غیر خطی [[27]](#footnote-27) در N بعد مي باشد که N بيانگر تعداد نمونه در داده هاي آموزش مي باشد. حل مساله استاندارد QP نياز به انجام عمليات روي ماتريس ها مي باشد همچنين زمان محاسباتي بالايي براي انجام دارد به همين دليل براي تعداد داده هاي زياد استفاده از آن عملي نمي باشد.

با استفاده از روش بهینه سازی کمینه ترتیبی [[28]](#footnote-28) مي توان مسئله ریزی غیر خطی براي ماشین بردار پشتیبان را با سرعت بيشتر و بدون ذخيره سازي اضافي ماتريس ها حل کرد اين روش در [26] توزيع داده شده است.

روش آموزش ماشین بردار پشتیبان نهايتا به بهينه سراسري مي رسد و در بهينه محلي گير نمي کند. همچنین این روش يک روش باينري مي باشد بنابراين در مسائل دسته بندي چند کلاسه بايد آن را به چندين مساله ماشین بردار پشتیبان دودويي تبديل کرد [27].

* + 1. مقايسه روش ها

به صورت عمومي روش هاي ماشین بردار پشتیبان و شبکه هاي عصبي در مواجعه با مسائل پيوسته و چندبعدي بهتر عمل مي کنند. براي مدل هاي شبکه هاي عصبي و ماشین بردار پشتیبان به تعداد زيادي نمونه در داده هاي آموزش براي رسيدن به خطاي کمتر در پيش بيني نياز مي باشد که در شبکه هاي بيزي اين مسئله به مراتب کمتر است.

الگوريتم هاي درخت تصميم نمي توانند براي مسائلي که به تقسيم بندي قطري نياز دارند به خوبي عمل کنند. در مسائلي که داده ها رابطه غير خطي با خروجي دارند روش هاي شبکه هاي عصبي و ماشین بردار پشتیبان به خوبي عمل مي کنند.

اگرچه زمان آموزش بر اساس نوع کاربرد و ديتاست متفاوت مي باشد. متخصصان عموما بر تقسيم بندي کلاس هاي اصلي الگوريتم هاي يادگيري توافق دارند. به عنوان مثال الگوريتم هاي تنبل داراي زمان آموزش صفر مي باشند به اين دليل که به سادگي داده هاي ورودي آموزش را ذخيره سازي مي کنند. روش هاي بیزی ساده [[29]](#footnote-29) به سرعت آموزش مي بينند به اين دليل که با يک گذر از داده ها و شمارش فراواني داده ها در نمونه هاي گسسته و يا محاسبه تابع چگالي نرمال در نمونه هاي پيوسته آموزش انجام مي شود.

درخت هاي تصميم تک متغيره نيز به سريع بودن در آموزش مشهور مي باشند و داراي مرتبه پايين تر نسبت به شبکه هاي عصبي و ماشین بردار پشتیبان مي باشند.

بیزی ساده به فضاي ذخيره سازي کمي هم در فرآيند آموزش و هم در فرآيند دسته بندي نياز دارند که به مينموم حافظه براي ذخيره سازي احتمالات پيشين و شرطي نياز دارند. الگوريتم k نزدیک ترین همسایه به فضاي ذخيره سازي زيادي براي فاز آموزش و حداقل همان مقدار فضا براي اجراي فاز دسته بندي دارند.

براي تمامي روش هاي يادگيري غير تنبل ميزان فضاي مورد نياز اجرا معمولا کمتر از فضاي مورد نياز براي آموزش مي باشد به اين دليل که فضاي اجرا خلاصه اي از داده هاي فضاي آموزش است.

بنابراين تعداد مدل يا زمان اجراي پارامتر هايي هستند که مي تواند با روش استفاده کاربر آن تنظيم شود و تنظيم پارامترها مي تواند تاثير مستقيم در کارايي الگوريتم داشته باشد. براي کاربري با تخصص کمتر در داده کاوي و يادگيري ماشين استفاده از الگوريتمي با تعداد پارامترهاي کمتر راحت تر و مناسب تر مي باشد ولي کاربري با داشتن دانش نسبت به داده کاوي با استفاده از الگوريتم هايي که پارامترهاي بيشتري دارند امکان مي دهد که کنترل بيشتري رو پروسه داده کاوي داشته باشد. ماشین بردار پشتیبان و شبکه هاي عصبي مصنوعي پارامترهاي بيشتري نسبت به بقيه تکنيک ها دارند.

الگوريتم k نزدیک ترین همسایه به دليل شيوه کاري که دارد نسبت به ويژگي هاي نامربوط بسيار حساس مي باشد همچنين وجود ويژگي هاي نامرتبط مي تواند پروسه آموزش شبکه هاي عصبي را ناموثر کند.

تفسير الگوريتم هاي مبتني بر منطق بر خلاف الگوريتم هاي شبکه هاي عصبي، ماشین بردار پشتیبان و k نزدیک ترین همسایه ساده تر مي باشد. هيچکدام از الگوريتم روي تمامي ديتاست ها به بقيه برتري ندارد. وقتي با اين مساله که کدام الگوريتم براي مسئله دسته بندي ما مناسب است رو به رو شديم، ساده ترين راه حل تخمين دقت الگوريتم هاي مختلف روي داده ها و انتخاب بهترين الگوريتم از نظر ميزان دقت است.

مفهوم ترکيب دسته بندی کننده ها با يکديگر مسيري جديد براي بهبود دسته بندی کننده در پيش روي ما قرار مي دهد. هدف ترکيب دسته بندی کننده ها ايجاد سيستم هايي با دقت بيشتر در نتايج مي باشد.

* 1. يادگيري بدون نظارت

گروه ديگري از الگوريتم هاي استاندارد يادگيري ماشين الگوريتم هاي يادگيري بدون نظارت مي باشند. ورودي اين الگوريتم ها تنها داده هاي خام (بدون برچسب) مي باشند. به عبارت ديگر ناظری براي يادگيري الگوريتم ها وجود ندارد.

يادگيري بدون نظارت مفهوم عميقي است که مي توان از ديدگاه هاي مختلف مانند از ديد روانشناسان و علوم شناختي و بررسي شود. اغلب از اين روش را يادگيري بدون معلم ياد مي شود که به اين معني است که انسان، حيوان و يا ماشين فقط با مشاهده محيط و بدون هيچ گونه آموزش رفتاري را پيش بگيرد. نتيجه آموزش بدون نظارت توصيفي جديد از داده هاي مشاهده شده است. محبوب ترين و نماينده اصلي اين نوع يادگيري الگوريتم هاي خوشه بندي مي باشد. در ادامه به بررسی روش های مهم کلاسترینگ پرداخته خواهد شد.

* + 1. خوشه بندی

با وجود بررسي هايي که از الگوريتم هاي خوشه بندي در منابع [28]، [29]، [30] و [31] در دامنه های مختلفي مانند يادگيري ماشين، داده کاوي، بازيابي اطلاعات، تشخيص الگو، بايو انفورماتيک و هستي شناسي شده است.

براي کاربران انتخاب الگوريتمي براي ديتاست هاي بزرگي که دارند به دلايل زير مشکل مي باشد و اين به دليل محدوديت هايي که در بررسي هاي انجامي وجود دارد است :

1- خصوصيات الگوريتم ها به خوبي مطالعه نشده است.

2- الگوريتم هاي جديد زيادي در اين شاخه ايجاد شده که در بررسي ها به خوبي مطالعه نشده اند.

3- مقايسه عملي خوبي ميان اين الگوريتم ها انجام نشده است.

در بخش بعد به بررسي انواع دسته بندي هاي روش هاي خوشه بندي خواهيم پرداخت.

* + 1. دسته بندي الگوريتم هاي خوشه بندي

در اين بخش به دسته بندي الگوريتم هاي خوشه بندي پرداخته شده است. اين دسته بندي انجام شده بر اساس ديدگاه يک طراح الگوريتم با توجه به جزئيات تکنيکي در فرآيند خوشه بندي است [32]. دسته بندي ارائه شده به صورت زير است :

* + - 1. **افراز [[30]](#footnote-30)**

در اين روش کلاستر هاي اوليه بلافاصله تشکيل شده و سپس در مراحل بعد تغيير مي کنند. به عبارت ديگر الگوريتم هاي مبتني بر پارتيشن، داده ها به تعدادي پارتيشن تقسيم مي کنند. هر پارتيشن بيانگر يکي از خوشه ها مي باشد که اين خوشه ها بايد شروط زير را برقرار نمايند :

1. هر خوشه حداقل بايد شامل يک عضو باشد.
2. هر عضو بايد دقيقا متعلق به يک خوشه باشد.

به عنوان مثال در الگوريتم k-means يک مرکز خوشه شامل ميانگين حسابي تمامي نقاط آن خوشه مي باشد. در الگوريتم k-medoids نقطه نزديک به مرکز هر خوشه نماينده آن خوشه مي باشد.

* + - 1. **سلسله مراتبی [[31]](#footnote-31)**

داده ها بر اساس ميانگين نزديکي به صورت سلسله مراتبي دسته بندي مي شوند. میزان نزديکی ميان داده ها توسط نودهاي مياني تعيين مي شوند. در واقع کل ديتاست توسط يک ساختار درختي دسته بندي مي شود که هر گره برگ بيانگر يکي از عناصر داده مي باشد.

خوشه بندي سلسله مراتبي مي تواند به صورت تراکمي ( پايين به بالا ) و يا تقسيم ( بالا به پايين ) انجام شود. روش تراکمي خوشه بندي را با قرار دادن هر عنصر داده در يک خوشه آغاز مي کند و به صورت بازگشتي دو يا تعداد بيشتري از خوشه ها را با هم ادغام مي کند.

در روش تقسيم خوشه بندي با يک خوشه که شامل کليه داده ها مي باشد آغاز شده و به صورت بازگشتي در هر مرحله به تقسيم کلاستر به تعداد کلاستر بيشتر مي پردازد.

اين مراحل به صورت بازگشتي تا رسيدن به يک شرايط پايان مانند رسيدن به تعداد مشخص خوشه ادامه پيدا مي کند. مشکل اصلي روش هاي سلسله مراتبي اين مي باشد که در صورتي که يک مرحله ( ادغام يا جداسازي ) انجام بگيرد نمي توان آن را برگرداند.

* + - 1. **مبتنی بر چگالی [[32]](#footnote-32)**

در روش هاي مبتني بر چگالي داده ها بر اساس چگالي داده ها در ناحيه اي که قرار دارند، ميزان اتصال و مرز آن ها با ديگر داده ها خوشه بندي مي شوند.

اين الگوريتم ها بيشتر سعي در تشکيل خوشه در بين نزديک ترين همسايه ها با هم دارند. يک خوشه، به عنوان يک جزء چگال که در مسيري که چگالي بيشتر است رشد کرده مشخص مي شود. به همين دليلي الگوريتم هاي خوشه بندي مبتني بر چگالي توانايي پيدا کردن خوشه هايي با اشکال غير هندسي را دارند.

همچنين ذات اين الگوريتم ها حفاظت در برابر داده هاي پرت را خود انجام مي دهد. بنابراين چگالي کلي يک نقطه براي تعيين توابعي از ديتاست که روي نقطه اي خاص در آن تاثير دارد آناليز مي شود.

* + - 1. **مبتنی بر گرید**

در اين روش فضاي داده تقسیم بندی می شود. هدف اصلي اين روش پردازش سريع مي باشد زيرا اين روش تنها يک بار ديتاست را براي محاسبات گريد پيمايش مي کند. اين روش مستقل از تعداد داده هاي ديتاسيت آن را گريد بندي مي کند و براساس گريد بندي ايجاد شده خوشه بندي را به جاي داده ها روي گريد اعمال مي کند.

کارايي روش گريدبندي بستگي به اندازه گريد دارد که معمولا اين اندازه کوچک تر از سايز داده هاي ديتاست است. اگرچه در توزيع هاي مختلف تقسيم بندي داده به صورت يکنواخت براي رسيدن به کيفيت خوشه بندي مناسب و يا زمان مناسب متناسب نمي باشد.

روش هاي Wave Cluster و STING [[33]](#footnote-33) از نمونه هاي روش هاي مبتني بر گريد براي خوشه بندي مي باشند.

* + - 1. **مبتنی بر مدل**

اين روش ها سعي در بهينه سازي ميزان سازگاري داده هاي ورودي و مدل هاي رياضي از پيش تعيين شده دارند. اين روش ها مبتني بر اين فرضيه مي باشد که داده هاي توليد شده بر اساس توزيع هاي آماري يا ترکيبي از اين توزيعات توليد شده اند. همچنين اين روش ها به راهي براي تشخيص خودکار کلاستر ها بر اساس مباني آمار استاندارد منجر مي شود که مي تواند داده هاي پرت و نويز را تحمل کرده و به يک روش قوي براي خوشه بندي برسد.

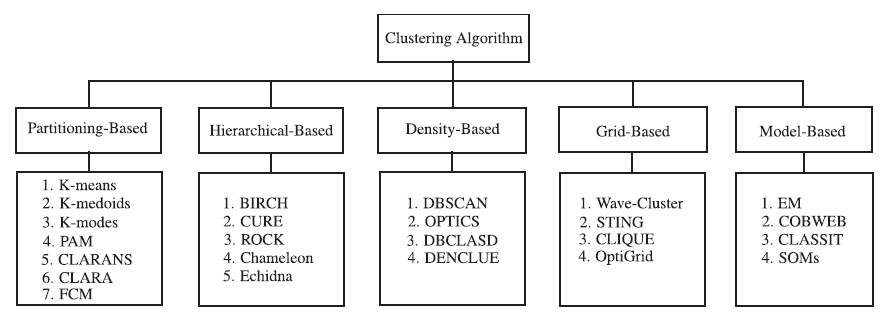
دو روش اصلي براي خوشه بندي مبتني بر مدل وجود دارد :

1- آمار

2- شبکه هاي عصبي مصنوعي

روش هاي آماري از احتمال براي اندازه گيري مفاهيم و خوشه بندي استفاده مي کند. شبکه هاي عصبي مصنوعي از از مجموعه از ورودي و خروجي هاي متصل به هم تشکيل شده است که هر کدام از اتصالات داراي وزني مرتبط با آن مي باشد. شبکه هاي عصبي چندين خصيصه دارند که باعث محبوبيت آن ها براي خوشه بندي شده است. مورد اول اينکه شبکه هاي عصبي مصنوعي ذاتا ساختاري موازي و توزيع شده دارند و مورد دوم اينکه شبکه هاي عصبي با تغيير در وزن هايش يادگيري را انجام مي دهند و به همين دليل مي توانند بهترين سازگاري با داده ها را ايجاد کنند. سوم داده هايي که توسط شبکه عصبي استفاده مي شوند وکتورهاي عددي مي باشند و نيازمند کمي بودن ويژگي هاي ورودي ها می باشند.

شکل 4 دسته بندي روش هاي ذکر شده را در قالب درخت توصيف مي کند [32].



شکل 1-1 : دسته بندی روش های Clustering

* + 1. الگوريتم هاي خوشه بندی

در این بخش به بررسی روش های معروف تر دسته بندی ارائه شده در بخش قبل پرداخته شده است.

* + - 1. الگوریتم FCM [[34]](#footnote-34)

FCM [33] الگوريتمي بر پايه منطق فازي و الگوريتم k-means براي خوشه بندي مي باشد. روش FCM يک روش خوشه بندي نرم مي باشد که هر عنصر داده با يک ميزان تعلق جدا به هر یک از خوشه ها متعلق مي باشد. بنابراين يک داده مي تواند با درجات متفاوت متعلق به چندين خوشه باشد.

اين روش ابتدا مرکز هر خوشه را مشخص شده و درگام بعد درجه عضويت براي هر داده نسبت به هر خوشه را محاسبه مي کند. الگوريتم FCM ميزان واريانس بين خوشه ها را مينيمم مي کند. این روش مشکل اصلي روش k-means را نيز به ارث برده است، به اين معني که خوشه بندي به صورت محلي عمل مي کند و خوشه هاي نهايي کاملا وابسته به وزن انتخابي براي شروع الگوريتم می باشد.

الگوريتم FCM همانند k-means در چند تکرار مرکز خوشه ها را جست و جو کرده و عضويت هر کدام از عناصر داده در خوشه ها را به روز مي کند. برخلاف روش k-means که عضويت به صورت سخت انتخاب مي شود و هر داده فقط متعلق به يک خوشه مي باشد در روش FCM عضويت هر داده در هر کلاستر را با عددي بين 0 و 1 نشان داده می شود، اين عدد بيانگر ميزان احتمال تعلق داده به آن خوشه مي باشد.

البته طبق يکي از قوانين منطق فازي مجموع تعلقات هر داده به کليه خوشه ها برابر يک بايد باشد.

* + - 1. الگوریتم BIRCH [[35]](#footnote-35)

الگوريتم BIRCH [34] يک ساختار درختي به نام درخت ويژگي خوشه بندي [[36]](#footnote-36) ايجاد مي کند. درخت ویژگی خوشه بندی مي تواند در يک روش افزايشي و پويا با اسکن ديتاست ايجاد شود. بنابراين اين روش از ابتداي فرآيند به کل داده ها نياز ندارد. اين روش دو فاز اصلي دارد :

1- ابتدا ديتاست براي ايجاد يک درخت در حافظه اسکن مي شود.

2-الگوريتم براي خوشه بندي نودهاي برگ اعمال مي شود.

درخت ویژگی خوشه بندی درختي متوازن مي باشد که بر اساس دو پارامتر فاکتور انشعاب B و حد آستانه T کار مي کند. وقتي که يک داده را مي خواهيم خوشه بندي کنيم درخت از ريشه پيمايش شده و نزديک ترين نود در هر سطح انتخاب مي شود. اگر نزديک ترين برگ کلاستر براي داده پيدا شد تستي براي اطمينان از تعلق داده به کلاستر کانديد انجام مي شود. در صورت عدم تعلق داده به کلاستر، کلاستري جديد با قطري بزرگتر ايجاد مي شود.

اين روش به صورت عمومي توانايي ايجاد يک خوشه بندي مناسب با يک بار پيمايش ديتاست را داشته و با چند مرحله تکرار مي تواند کيفيت آن را بهبود دهد همچنین توانايي کنترل مناسب داده هاي پرت و نويز را نيز دارد.

البته اين روش به دليل استفاده از مفاهيم قطر و شعاع در مواردي که خوشه ها به صورت کروي نباشد به خوبي کار نمي کند همچنين اين روش به اريب بودن داده ها حساس بوده و بر اساس ترتيب مختلف ورودي در داده هاي يکسان مي تواند خوشه هاي متفاوت توليد کند.

* + - 1. الگوریتم DENCLUE [[37]](#footnote-37)

الگوريتم DENCLUE [35] بر اساس مدل هاي آماري با توجه به ميزان توابع تاثير تمام داده ها خوشه بندي را انجام مي دهد. تابع نفوذ را مي توان به عنوان تابعي که ميزان تاثير داده ها بر همسايگانش را تعيين مي کند در نظر گرفت و نواحي که چگالي را به خود جذب مي کنند را به عنوان خوشه در نظر گرفت.

در اين الگوريتم خوشه هايي با شکل هاي نامنظم و غير هندسي مي توان به صورت يک معادله ساده در کرنل تابع چگالي مشخص شوند.

تعيين مناسب مقادير پارامترهاي این الگوریتم تاثير زيادي در خوشه هاي بدست آمده دارد. اين روش در مقايسه با بقيه روش هاي خوشه بندي مزايايي دارد :

1- اين روش يک پايه رياضي محکمي داشته و ميتواند روش هاي خوشه بندي ديگر مانند روش هاي تقسيم و سلسله مراتبي را پياده سازي کند.

2- اين روش براي ديتاست هايي با داده هاي نويز زياد مناسب است.

3- اين روش به ما امکان پيدا کردن خوشه هايي با اشکال غيرهندسي در در داده هايي با ابعاد بالا مي دهد.

4- اين روش از سلول هاي گريد براي نگه داري اطلاعات در رابطه با سلول هاي حاوي داده استفاده مي کند. اين سلول ها را در ساختاري درختي نگه داري مي کند بنابراين نسبت به بقيه الگوريتم هاي اين دسته سريع تر مي باشد.

کليه اين ويژگي ها باعث ايجاد کلاسترهاي خوب در ديتاست هاي بزرگ داراي نويز زياد مي شود..

* + - 1. الگوریتم OptiGrid [[38]](#footnote-38)

الگوريتم OptiGrid [36] براي بدست آوردن تقسیم بندی بهينه طراحي شده است. اين روش با ايجاد بهترين برش فوق صفحه از ميان مجموعه طرح هاي انتخابي سعي در ايجاد بهترين گريد دارد. اين طرح ها براي پيدا کردن بهترين برش ها استفاده مي شوند. هر برش به گونه اي انتخاب مي شود که داراي کمترين نقطه چگالي باشد و بتوان ناحيه با چگالي بيشتر را به دو قسمت تقسيم کند. بعد از هر مرحله از ايجاد گريد چند بعدي الگوريتم با استفاده از تابع چگالي خوشه ها را پيدا مي کند در مرحله بعد الگوريتم به صورت بازگشتي به خوشه ها اعمال مي شود. در هر مرحله بازگشتي الگوريتم تنها داده هاي گريد ها با چگالي بالا را نگه داري و افزايش مي دهد.

اين روش براي خوشه بندي داده ها با ابعاد بالا بسيار کارا مي باشد البته اين الگوريتم در پيدا کردن خوشه هايي با ابعاد کم در يک فضا با ابعاد زياد بسيار ضعيف عمل مي کند. عملکرد اين الگوريتم شديدا به پارامترهاي انتخابي وابسته است و نميتواند مديريت خوبي در گريد هايي با سايز بيشتر از حافظه داشته باشد.

* + - 1. الگوریتم EM [[39]](#footnote-39)

الگوريتم EM [37] براي تخمين بيشترين شباهت پارامترهاي يک مدل آماري در موقعيت هاي مختلف طراحي شده است که مي توان به موقعيت هايي نمی توان مستقيما معادلات را حل کرد اشاره نمود.

الگوريتم EM در چند تکرار مقادير مجهول پارامترهاي مدل تخمين مي زند در هر تکرار دو مرحله E و M اجرا می شود.

در مرحله E مقادیر مدل جاري براي ارزيابي توزيع آخرين مقادير متغير ها استفاده مي شود و داده ها بر اساس توزيع آخرين مقادير متغير ها در کلاسترها تقسیم می شوند. در مرحله M مقادير متغيرها بر اساس قانون ماکسيمم شباهت پیش بینی می شوند.

الگوريتم EM پيدا کردن ماکسيموم محلي را براي تخمين مقادير ضمانت مي کند. مهمترين معايب اين روش نياز به ماتريس غيرمنفرد کواریانس، حساسيت به انتخاب پارامترهاي ورودي، امکان همگرايي به يک بهينه محلي و نرخ همگرايي آهسته مي باشد.

* + 1. بررسي معيار هاي سنجش روش هاي خوشه بندي

در ارزیابی روش هاي کلاسترینگ بايد شرايط مشخصي براي ارزيابي ضعف ها و نقات قوت هر الگوريتم با توجه به ويژگي هاي اصلي داده که شامل سرعت، ميزان و تنوع داده ها مي باشد تعيين کنيم. در اين قسمت اين خصيصه ها تعريف شده و خصوصيات کليدي هر خصيصه بررسي شده است.

* + - 1. نوع ديتاست

بيشتر الگوريتم هاي قديمي خوشه بندي براي تمرکز روي داده هاي عددي و يا داده هاي دسته بندي شده طراحي شده بودند. داده هاي جمع آوري شده در دنياي واقعي اغلب داده ها شامل هم داده هاي عددي و هم داده هاي دسته بندي شده اند و استفاده از روش هاي قديمي خوشه بندي به طور مستقيم روي اين داده ها مشکل مي باشد و الگوريتم هاي خوشه بندي روي داده هايي که خالص عددي يا دسته بندي شده اند به صورت موثر کار ميکنند و اغلب روي ترکيب اين داده ها کارايي مناسبي ندارند.

* + - 1. سايز ديتاست

سايز ديتاست تاثير مهمي در کيفيت خوشه بندي دارد. برخي از روش هاي خوشه بندي تاثير موثرتري نسبت به بقيه در خوشه بندي داده هايي با سايز کوچک و يا سايز بزرگ دارند.

* + - 1. پارامترهاي ورودی

يک ويژگي مهم براي خوشه بندي تعداد کمتر پارامترهاي ويژگي مي باشد. از آنجا که تعداد زياد پارامترها ممکن است بر کيفيت خوشه بندي تاثيرگذار مي باشد. به اين دليل که خوشه بندي وابسته به اين مقادير ورودي است.

* + - 1. مديريت داده هاي پرت و نويز

يک الگوريتم موفق اغلب توانايي مديريت داده هاي نويز و داده هاي پرت را دارد. از آنجا که داده هاي ورودي اغلب داده هاي دنياي واقعي مي باشند خالص نبوده و همواره با نويز همراه مي باشند و نويز نيز خوشه بندي داده ها را مشکل مي کند و نتايج الگوريتم را تحت تاثير قرار مي دهد.

* + - 1. پيچيدگي زماني

بيشتر روش هاي خوشه بندي بايد چندين مرتبه اجرا شوند تا کيفيت خوشه هاي ارائه شده توسط آن ها افزايش يابد به همين دليل اين پروسه زمانبر مي باشد و مي تواند براي کاربردهاي مرتبط با داده های حجیم غير عملي باشد.

* + - 1. پایداری

يکي از ويژگي هاي مهم هر الگوريتم هاي خوشه بندي قابليت توليد يک خوشه بندي يکسان بدون توجه به ترتيب و الگويي که داده ها به آن داده شده اند است.

* + - 1. چگونگی برخورد با داده های حجم بالا

اين ويژگي يکي از مهمترين ويژگي هاي در خوشه بندي مي باشد به اين دليل که در بيشتر کاربردها نيازمند به آناليز داده هايي با ابعاد ويژگي هاي زياد مي باشيم. به عنوان مثال اسناد متني مي توانند داراي هزاران واژه يا کلمات کليدي باشند و اين مسائله مي تواند چالش بر انگيز باشد همچنين بسياري از اين ابعاد مي توانند نامرتبط باشند. با افزايش ابعاد داده ها مي توانند به شدت خلوت شوند و تعريف فاصله بين داده ها مشکل مي شود و داده ها داراي چگالي کم مي شوند.

* + - 1. شکل خوشه

يک الگوريتم خوشه بندي خوب بايد بتواند داده هاي واقعي و تنوع آن ها را مديريت کند و بتواند در صورت نياز کلاسترهايي با اشکال غير هندسي ايجاد کند.

* 1. نتیجه گیری

در این فصل مفهوم داده کاوی شرح داده شد.

در فصل بعد تحقیقات اخیر انجام شده در زمینه داده کاوی پزشکی مورد بررسی قرارخواهند گرفت.

فصل سوم: مروری بر تحقیقات اخیر انجام شده

3-1- مقدمه

در فصل قبل به چالش­های که در داده کاوی وجود دارد پرداخته شد. در این فصل به تحقیق­های جدیدی که در این زمینه انجام داده شده است، پرداخته خواهد شد.

3-2- مروری بر ادبیات موضوع

در تحقیق انجام شده توسط استیل و همکارانش در [38] خوشه بندی مشکلات و علائم بیماران سرطانی و غیرسرطانی تحت مراقبت های تسکینی انجام می شود. این تحقیق، اولین تحقیق انجام شده برای کشف خوشه های مشکلات و علائم بیماران غیرسرطانی در مراقبت های تسکینی می باشد. به طور کلی، شناسایی خوشه های مشکلات و علائم بیماران می تواند به پزشکان در تشخیص آن ها و کشف درمان مناسب کمک کند. خوشه بندی علائم برای کشف رابطه بین آن ها انجام می شود تا پزشکان بتوانند به کمک آن ها استراتژی درمانی استاندارد برای هر علامت و خوشه هایی از آن ها تعیین کنند. از روش سلسله مراتبی تجمیعی [[40]](#footnote-40) برای خوشه بندی استفاده می شود. از داده های بیماران بستری مربوط به مراقبت های تسکینی و بیمارستانی آلمان استفاده می شود. از 16 مشکل و علامت برای خوشه بندی استفاده می شود. تعداد بیماران سرطانی برابر 6181 و بیماران غیرسرطانی برابر 560 نفر می باشد. مشکلات و علائم بیماران سرطانی و همچنین مشکلات و علائم بیماران غیرسرطانی به 5 خوشه تقسیم می شوند.

در تحقیق انجام شده توسط ماسکیاری و همکارانش در [39] الگوریتم خوشه بند M-CLUBS [[41]](#footnote-41) برای داده های ریزآرایه ارائه شده است. این الگوریتم دقت بالاتری نسبت به الگوریتم های سلسله مراتبی و پیچیدگی زمانی کمتری نسبت به الگوریتم های افراز دارد. از دیتاست های مختلف ریزآرایه برای بررسی نتایج استفاده می شود. ویژگی های اصلی الگوریتم عبارتند از : استفاده از گریدهای پویا (برخلاف الگوریتم هایی مثل STING) اصلاح خطاهای مراحل قبل، تشخیص خوشه های غیرکروی شکل (برخلاف الگوریتم هایی k-means و BIRCH) و عدم نیاز به پارامترهای ورودی کاربر. از چهار دیتاست ریزآرایه برای مقایسه نتایج استفاده شده است. نتایج بدست آمده نشان می دهد ارائه شده برای داده های ریزآرایه نسبت به الگوریتم هایی نظیر OPTICS، BIRCH، k-means++، SMART، DIANA، UPGMA، NN، FN، SiMM-TS دارای خطاهای کمتری می باشد.

در تحقیق انجام شده توسط چن و همکارانش در [40] یک مدل ترکیبی هوشمند که از تحلیل خوشه و انتخاب ویژگی برای تحلیل تشخیص های سرطان سینه استفاده می کند ارائه شده است. در روش ارائه شده قبل از انجام خوشه بندی، انتخاب ویژگی انجام می شود تا کیفیت خوشه بندی افزایش یابد. از یک روش انتخاب ویژگی به نام IBNF استفاده می کند که یک روش مبتنی بر فیلتر [[42]](#footnote-42) می باشد و آن را با سه روش مبتنی بر فیلتر دیگر و چهار روش مبتی بر بسته بندی مقایسه می کند. در اولین آزمایش روش IBNF با سه روش مبتنی بر فیلتر و در آزمایش دوم با چهار روش مبتی بر بسته بندی مقایسه می شود. از دو دیتاست متفاوت مخزن UCI [[43]](#footnote-43) در دو آزمایش انجام شده استفاده می کند. در آزمایش اول تعداد ویژگی ها برابر 32 و تعداد نمونه ها برابر 569 می باشد. در آزمایش اول تعداد ویژگی ها برابر 11 و تعداد نمونه ها برابر 683 می باشد. از معیار R-Squared برای مقایسه روش های تست شده استفاده می شود. نتایج بدست آمده نشان می دهد که روش IBNF دارای عملکرد بهتری نسبت به سایر روش ها می باشد.

در تحقیق انجام شده توسط چائو و همکارانش در [41] پیش بینی بقای سرطان سینه با استفاده از سه روش رگرسیون منطقی [[44]](#footnote-44)، ماشین بردار پشتیبان [[45]](#footnote-45) و درخت تصمیم [[46]](#footnote-46) انجام شده است. برای ساخت درخت تصمیم از الگوریتم C5.0 استفاده شده است. نتایج بدست آمده نشان می دهد این سه روش از صحت بالایی در پیش بینی بقای سرطان سینه برخوردار هستند. روش ماشین بردار پشتیبان دارای بیشترین صحت در بین این سه روش می باشد. از روش 10-Fold Cross-Validation برای تشخیص مدل استفاده شده است. از داده های بیمارستان تایوان شامل 1340 نمونه و 8 ویژگی استفاده شده است. از معیار صحت برای مقایسه نتایج استفاده شده است. صحت بدست آمده برای درخت تصمیم C5.0 برابر 95/93 %، برای رگرسیون منطقی برابر 1/95 % و برای ماشین بردار پشتیبان برابر 22/95 % می باشد. نتایج بدست آمده نشان می دهد ماشین بردار پشتیبان دارای بهترین صحت در بین سه روش بررسی شده می باشد.

در تحقیق انجام شده توسط مونت و همکارانش در [42] از رگرسیون منطقی در تحلیل داده های ریزآرایه برای پیش بینی سرطان ریه و سرطان سینه استفاده شده است. برای بررسی میزان صحت روش رگرسیون از منحنی ROC [[47]](#footnote-47) استفاده شده است. در اولین مرحله از تحلیل رگرسیون نمونه های مناسب از بین داده ها انتخاب می شوند. برای انتخاب ژن های مناسب از روش آماری ANOVA [[48]](#footnote-48) استفاده شده است. از روش Leave-One-Out و Four Group Split برای تشخیص مدل استفاده شده است. از دیتاست NCBI GEO [[49]](#footnote-49) برای هر دو نوع سرطان استفاده شده است. نتایج بدست آمده نشان می دهد با افزایش تعداد ژن ها، صحت پیش بینی افزایش می یابد.

در تحقیق انجام شده توسط بریمز و همکارانش در [43] از درخت دسته بندی و رگرسیون [[50]](#footnote-50) برای توسعه مدل پیش بینی کننده سرطان ریه به منظور بررسی میزان احتمال بقای بیمار استفاده شده است. بیماران در دو 4 گروه دسته بندی می شوند ( با استفاده از درخت دسته بندی و رگرسیون) و سپس روش Cox Proportional Hazard Regression که یک روش آماری است و بقای بیمار را با توجه به متغیرهای مروبط به آن پیش بینی می کند، برای پیش بینی بقای بیمار استفاده می شود. از 29 متغیر مستقل استفاده شده است و مدت زمان بقای 18 ماهه در نظر گرفته شده است. برای پارتیشن کردن یک نود در درخت تصمیم از معیار Chi Square استفاده می شود. محاسبات آماری و تغییرات داده ها با استفاده از نرم افزار SPSS انجام شده است. از داده های بیمارستان مرکز سرطان استرالیا که شامل 29 متغیر و 482 نمونه می باشد استفاده شده است. نتایج بدست آمده نشان می دهد مقدار حساسیت [[51]](#footnote-51) برابر 5/94 %، مقدار وضوح [[52]](#footnote-52) برابر 2/38 %، مقدار C-Static برابر 1/76 % و مقدار پیش بینی کننده مثبت [[53]](#footnote-53) برابر 76 % می باشد.

به کشف میزان سرطان و مکان آن در بدن بیمار مرحله بندی گفته می شود. مرحله بندی قبل از شروع درمان انجام می شود. دو نوع مرحله بندی وجود دارد : مرحله بندی پاتالوژی و کلینیکی. مرحله بندی کلینیکی از روی آزمایشات مثل X-Ray و CT Scan بدست می آید. مرحله بندی پاتالوژی از روی آزمایشات و اطلاعاتی حین جراحی بدست می آید. مزایای این روش عبارتند از حفظ منابع پزشکی، افزایش تاثیر درمان، افزایش تشخیص و پیشگیری زودرس و.... در تحقیق انجام شده توسط یانگ و همکارانش در [44] استنتاج گزارش پاتولوژی بدون عمل جراحی بوسیله اطلاعات کلینیکی که از کاوش مشارکتی [[54]](#footnote-54)برای انجام این کار استفاده می کند مورد توجه قرار گرفته است. از تکنیک های داده کاوی برای پیدا کردن همبستگی [[55]](#footnote-55) بین اطلاعات کلینیکی و گزارش پاتولوژی برای تشخیص سرطان ریه استفاده شده است. از الگوریتم Apriori استفاده می شود تا میزان محاسبات کاهش یابد.

در تحقیق انجام شده توسط زیبا و همکارانش در [45] پش بینی وضعیت بیماران سرطان ریه بعد از عمل جراحی با استفاده از ماشین بردار پشتیبان تلفیقی بر روی داده های نامتوازن [[56]](#footnote-56) انجام شده است. از روش پیشنهاد شده برای استخراج قوانین از داده های سرطان ریه استفاده می شود. از روش Oracle-Based برای استخراج قوانین استفاده شده است. از Information Gain برای کاهش ابعاد استفاده شده است. از 44 دیتاست موجود در ابزار KEEL [[57]](#footnote-57) استفاده شده است.

در تحقیق انجام شده توسط شریف الدین و همکارانش در [46] ترکیبی از روش های استدلال مبتنی برمورد [[58]](#footnote-58) و استدلال مبتنی بر قانون [[59]](#footnote-59) برای افزایش صحت دسته بندی داده های پزشکی نسبت به سیستم مبتنی بر مورد بر اساس بازیابی [[60]](#footnote-60) ارائه شده است. بعد از دسته بندی هر مورد، این روش گسترش می یابد و قوانین به صورت خودکار به روزرسانی می شوند. روش ارائه شده باعث افزایش صحت در داده های سرطان سینه و بیماری تیروئید می شود. از دو دیتاست مربوط به سرطان سینه که شامل داده های تصویری می باشد و بیماری تیروئید استفاده می شود. صحت بدست آمده برای بیماری تیروئید با روش مبتنی بر مورد بر اساس بازیابی برابر 49/93 % و برای روش پیشنهاد شده برابر 53/99 % می باشد. همچنین صحت بدست آمده برای سرطان سینه برابر در روش اول برابر 62/78 % و برای روش پیشنهاد شده برابر 33/99 % می باشد.

در تحقیق انجام شده توسط پینگ و همکارانش در [47] از داده های سری زمانی برای پیش بینی بازگشت سرطان کبد استفاده می شود. از یک استدلال مبتنی مورد چند معیاری [[61]](#footnote-61) برای ایجاد مدل پیش بینی کننده بازگشت سرطان کبد استفاده می شود. نتایج استدلال مبتنی بر مورد چند میعاری و تک معیاری [[62]](#footnote-62) ارزیابی و مقایسه می شوند و با توجه به نتایج بدست آمده استدلال مبتنی بر مورد چند معیاری کارایی بیشتری نسبت به تک مقداری دارد. از داده های بیمارانی که از روش RFA [[63]](#footnote-63) برای درمان سرطان کبد استفاده کرده اند استفاده می شود. داده های جمع آوری شده مربوط به بازه زمانی 180 روزه می باشند. صحت بدست آمده برای استدلال مبتنی بر مورد تک معیاری برابر 61 % و برای استدلال مبتنی بر مورد چند معیاری برابر 69 % می باشد که نشان می دهد استدلال مبتنی بر مورد دارای صحت بیشتری در پیش بینی سرطان کبد دارد.

در تحقیق انجام شده توسط منصوری و همکارانش در [48] یک روش انتخاب ویژگی فازی در تولید دسته بند هایی برای داده های حجم بالا ارائه شده است. از ویژگی های انتخاب شده در تولید قوانین دسته بند مبتنی بر قاعده فازی استفاده می شود. هریک از قوانین دسته بند مبتنی بر قانون فازی با یک روش پیشنهاد شده ارزیابی می شوند. روش پیشنهاد شده از درجه عضویت الگوهای مثبت و منفی در محاسبه میزان وابستگی ویژگی ها به کلاس های مختلف استفاده می کند. پیچیدگی محاسباتی این روش برابر O(nM) می باشد که n برابر تعداد ویژگی ها و M برابر تعداد کلاس ها می باشد. پیچیدگی محاسباتی دسته بند طراحی شده برابر O(MK) می باشد که در آن به ازای هر کلاس K قانون ارزیابی می شود.

در تحقیق انجام شده توسط مجید و همکارانش در [49] یک روش جدید برای پیش بینی سرطان سینه و روده بر روی داده های نامتوازن ارائه شده است. روش پیشنهاد شده شامل دو مرحله می باشد. مرحله اول پیش پردازش داده ها می باشد که در آن از تکنیک MTD [[64]](#footnote-64) برای افزایش تعداد نمونه هایی که در اقلیت هستند به منظور متوازن کردن دیتاست استفاده می شود و مرحله دوم پیش بینی داده ها می باشد از روش های یادگیری ماشین مانند ماشین بردار پشتیبان و k نزدیکترین همسایه [[65]](#footnote-65) و ایجاد مدل های ترکیبی MTD-SVM [[66]](#footnote-66) و MTD-KNN [[67]](#footnote-67) برای پیش بینی استفاده می شود. روش MTD-SVM دارای بهترین صحت و کمترین پیچیدگی زمانی در بین روش های بررسی شده می باشد. از سه دیتاست واقعی مربوط به سرطان سینه و روده استفاده شده است. بالاترین صحت پیش بینی روش MTD-SVM برابر 50/96 % می باشد.

در تحقیق انجام شده توسط ذولبنین و همکارانش در [50] تاثیر وجود بیماری های مزمن به طور همزمان در یک فرد بر روی پیش بینی زنده ماندن بررسی شده است. بیماری هایی که به طور همزمان در یک فرد وجود دارند بر روی تشخیص، معالجه، بقا و همچنین هزینه درمان تاثیر می گذارد. از 4 روش جنگل تصادفی [[68]](#footnote-68)، شبکه عصبی مصنوعی [[69]](#footnote-69)، درخت تصمیم و رگرسیون استفاده می شود. توجه به وجود همزمانی بیماری ها می تواند کیفیت سيستم پشتيبان تصميم [[70]](#footnote-70) را افزایش دهد. سرطان های ترکیب شده همبستگی بیشتری دارند و تعداد افرادی بیشتری را پوشش می دهند. از دیتاست SEER [[71]](#footnote-71) در این مقاله استفاده می شود. SEER یک دیتاست عمومی است که به صورت آنلاین قایل دستیابی می باشد. روش جنگل های تصادفی نتایج بهتری را نشان می دهد و با داده ای استفاده شده تناسب بیشتری دارد. مقادیر صحت، حساسیت و وضوح به برای روش جنگل تصادفی به ترتیب برابر 2/75 %، 6/65 % و 6/75 % می باشد.

در تحقیق انجام شده توسط کومار و همکارانش در [51] تست های مختلف آماری مبتنی بر MapReduce برای انتخاب ویژگی های مناسب استفاده شده است. بعد از مرحله انتخاب ویژگی از ماشین بردار مبدایی [[72]](#footnote-72) مبتنی بر MapReduce برای دسته بندی داده ریزآرایه استفاده شده است. الگوریتم های مربوطه بر روی فریم ورک هدوپ پیاده سازی شده اند. روش ذکر شد بر روی دیتاست های ریزآرایه مختلف با ابعاد مختلف تست می شود. طراحی مدل برای دسته بندی نمونه های دیتاست های مختلف در کلاس مربوط به آن ها انجام شده است. تست های آماری شامل ANOVA، Kruskal-Wallis و Friedman برای انتخاب ویژگی استفاده شده اند. از دیتاست NCBI GEO استفاده شده است. دیتاست NCBI GEO یک دیتاست عمومی می باشد که به صورت آنلاین قابل دستیابی می باشد. صحت بدست آمده برای برخی دیتاست ها از جمله دیتاست مربوط به سرطان خون برابر 100 % می باشد.

در تحقیق انجام شده توسط وانگ و همکارانش در [52] پیش بینی بقای بیماران سرطان مری با استفاده از سیستم استنتاج فازی [[73]](#footnote-73) انجام شده است. برای پردازش مقادیر ورودی از منطق فازی استفاده می شود و سپس از سیستم استنتاج فازی برای آموزش مدل استفاده می شود. معماری این سیستم دارای یک معماری 5 لایه از شبکه عصبی می باشد. از یک دیتاست واقعی مربوط به بیماران تایوانی شامل 271 نمونه استفاده شده است.

در تحقیق انجام شده توسط تی سای و همکارانش در [53] ژن های سرطان با استفاده از دسته بند های درخت تصمیم کشف می شوند. از نرم افزارهای STATISTICA و IPA برای انتخاب ویژگی و تحلیل مسیر [[74]](#footnote-74) (برای کشف تاثیر ژن های مختلف در پیشرفت سرطان) استفاده شده است. برای ساخت درخت تصمیم از الگوریتم های C&RT [[75]](#footnote-75) و CHAID [[76]](#footnote-76) استفاده شده است. داده های بیماران بیمارستان تایوان شامل 41 نمونه از بیماران سرطانی که در4 مرحله مختلف از سرطان قرار دارند. نتایج بدست آمده نشان می دهد به طور میانگین برابر 5/80 % می باشد.

در تحقیق انجام شده توسط تنها و همکارانش در [54] دسته بندی نیمه ناظر با استفاده از درخت تصمیم بر روی چند دیتاست مخزن UCI شامل سرطان سینه، کبد، هپاتیت و... انجام می شود. تغییرات مختلفی به درخت تصمیم داده می شود تا تخمین احتمال (تسبت کلاس اکثریت نمونه ها در برگ های درخت هرس شده) را به صورت دقیق تر انجام دهد. تغییرات انجام شده باعث افزایش کارایی داده های برچسب دار نخواهد بود و فقط برای داده های بدون برچسب در آموزش پیش خود [[77]](#footnote-77) مناسب است. از 14 دیتاست مخزن UCI که شامل داده های پزشکی و غیرپشکی می باشد استفاده شده است. از چندین روش دسته بندی برای مقایسه نتایج از جمله درخت تصمیم و بیز ساده استفاده شده است که روش پیشنهاد شده باعث افزایش صحت دسته بندی داده های سرطان سینه و بعضی داده های دیگر شده است. نتایج یادگیری باناظر برای درخت تصمیم و بیز ساده برابر 68 % و 5/72 % می باشد. نتایج یادگیری نیمه ناظر برای درخت تصمیم و بیز ساده برابر 12/70 % و 14/76 %می باشد.

دسته بندی مشارکتی [[78]](#footnote-78) ترکیبی از کاوش قوانین انجمنی و دسته بندی می باشد که در کاربردهای پزشکی مورد توجه قرار گرفته اند و از صحت بالایی برخوردار هستند. در تحقیق انجام شده توسط نانزینی و همکارانش در [55] الگوریتم CPAR [[79]](#footnote-79) اصلاح می شود تا صحت دسته بندی افزایش یابد. برای اصلاح الگوریتم CPAR، قوانین کیفی با استفاده از Gain Ratio تولید می شوند. قبل از اجرای الگوریتم CPAR، از روش های T-Test و Reduct Sets برای کاهش ابعاد استفاده می شود. براساس آزمایشات انجام شده، ترکیب CPAR-GR و T-Test بهترین صحت را ارائه می دهد. صحت این روش به طور میانگین برای داده های بررسی شده برابر 43/87 % می باشد.

در تحقیق انجام شده توسط گوش و همکارانش در [56] ترکیب منطق فازی و دسته بندی مشارکتی برای یافتن روابط بین ژن ها به منظور تشخیص حالت های سرطانی و غیرسرطانی انجام شده شده است. مراحل روش ترکیبی ارائه شده شامل تولید مجموعه های فازی از ژن ها، تولید آیتم های فازی، محاسبه پشتیبان فازی برای آیتم های فازی، محاسبه رابطه وابستگی بین دو آیتم فازی، تولید قوانین مشارکتی و تحلیل قوانین مشارکتی برای تشحیص حالت های سرطانی و غیرسرطانی می باشد. از 5 دیتاست ریزآرایه سرطان مربوط به دیتابیس GEO NCBI استفاده شده است. از معیار F-Score برای مقایسه نتایج استفاده شده است. داده های مربوط به سرطان ریه در روش پیشنهاد شده با F-Score 82 % دارای بهترین عملکرد می باشد.

در تحقیق انجام شده توسط کوئو و همکارانش در [57] یک شبکه دو مرحله ای عصبی فازی [[80]](#footnote-80) برای پیش بینی سرطان پروستات پیشنهاد شده است. پارامترهای تابع عضویت بوسیله تحلیل کلاستر بدست آمده است. برای کشف رابطه ورودی و خروجی، ترکیبی از الگوریتم های Opt-aiNET [[81]](#footnote-81) و PSO [[82]](#footnote-82) استفاده شده است. روش پیشنهاد شده پیش بینی سرطان پروستات را دقیق تر از روش های قبل انجام می دهد. برای پیدا کردن فاکتورهای مهم از تحلیل رگراسیون استفاده می شود و سپس آن ها را به شبکه عصبی فازی اعمال می کند. برای کلاستر کردن ویژگی ها از الگوریتم aiNet-K [[83]](#footnote-83) استفاده شده است. دیتاست استفاده شده در این مقاله از یک مرکز آموزشی درمانی در تایپه تهیه شده است. برای مقایسه نتایج از میانگین مربعات خطا [[84]](#footnote-84) استفاده شده است. برای هر آزمایش تکرار 500 درنظر گرفته شده است. میانگین مربعات خطا در سه بنچ مارک مختلف بررسی شده است. روش استفاده شده در این تحقیق دارای بهترین عملکرد نسبت به سایر روش های بررسی شده می باشد. مقدار میانگین مربعات خطا برای بنچ مارک Ackley Function به طور میانگین برابر 008/0، برای بنچ مارک Hartman Function به طور میانگین برابر 005/0 و برای بنچ مارک Mackey Glass Time Series به طور میانگین برابر 004/0 می باشد.

در تحقیق انجام شده توسط لو و همکارانش در [58] یک سیستم هوشمند برای تشخیص سرطان ریه با استفاده از الگوریتم ژنتیک به عنوان روش انتخاب ویژگی ارائه شده است. یک روش جدید برای انتخاب ویژگی با استفاده از الگوریتم ژنتیک [[85]](#footnote-85) ارائه شده است. از یک تابع برازندگی [[86]](#footnote-86) جدید برای ارزیابی ویژگی های کاندید استفاده شده است. از Mutual Information برای کشف وابستگی غیرخطی بین ویژگی ها استفاده شده است. پس از انجام انتخاب ویژگی، دسته بندی با روش های ماشین بردار پشتیبان، پس انتشار [[87]](#footnote-87) و k نزدیک ترین همسایه انجام می شود و با نتایج روش های استفاده از کل ویژگی ها، F-Score و انتخاب ویژگی مبتنی بر همبستگی [[88]](#footnote-88) مقایسه می شود. از داده های سرطان ریه مخرن UCI استفاده شده است. این دیتاست شامل 32 نمونه و 56 ویژگی برای سه نوع سرطان ریه می باشد. نتایج بدست آمده نشان می دهد ماشین بردار پشتیبان برای روش پیشنهاد شده دارای بهترین عملکرد می باشد. صحت این روش برابر 99 % می باشد.

3-3- نتیجه گیری

در این فصل به تحقیقات جدید انجام شده در زمینه داده کاوی پزشکی پرداخته شد. با بررسی کارهای انجام شده در این زمینه می توان دریافت که در سال های اخیر به روش های ترکیبی داده کاوی از جمله شبکه عصبی فازی، دسته بندی مشارکتی، دسته بندی تلفیقی و... توجه بیشتری شده است. برخی روش ها مانند ماشین بردار پشتیبان کاربرد بیشتری در داده کاوی پزشکی داشته و نتایج بهتری را نسبت به سایر روش ها ارائه می دهد

مراجع

مراجع

[1] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction To Data Mining*. Boston: Pearson Addison Wesley, 2005.

[2] S. Sumathi and S.N. Sivanandam: *Data Warehousing, Data Mining, and Olap*, Studies in Computational Intelligence (SCI) **29**, 21–73 (2006).

[3] T. G. Dietterich, “Machine-Learning Research,” *AI Mag.*, vol. 18, no. 4, p. 97, 1997.

[4] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, “A review of microarray datasets and applied feature selection methods,” *Inf. Sci. (Ny).*, vol. 282, pp. 111–135, 2014.

[5] S. Zhang, C. Zhang, and Q. Yang, “Data preparation for data mining,” *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 375–381, 2003.

[6] V. J. Hodge and J. Austin, “A Survey of Outlier Detection Methodoligies,” *Artif. Intell. Rev.*, vol. 22, no. 1969, pp. 85–126, 2004.

[7] G. Batista and M. C. Monard, “An analysis of four missing data treatment methods for supervised learning,” *Appl. Artif. Intell.*, vol. 17, no. Dm, pp. 519–533, 2003.

[8] L. Yu and H. Liu, “Efficient Feature Selection via Analysis of Relevance and Redundancy,” *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec. 2004.

[9] S. Markovitch and D. Rosenstein, “Feature generation using general constructor functions,” *Mach. Learn.*, pp. 59–98, 2002.

[10] S. K. Murthy, “Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey,” *Data Min. Knowl. Discov.*, vol. 2, no. 4, pp. 345–389, 1998.

[11] T. Elomaa, “Advances in Intelligent Data Analysis: Third International Symposium, IDA-99 Amsterdam, The Netherlands, August 9--11, 1999 Proceedings,” D. J. Hand, J. N. Kok, and M. R. Berthold, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 63–74.

[12] T. Lim, “A Comparison of Prediction Accuracy , Complexity , and Training Time of Thirty-three Old and New دسته بندی Algorithms,” vol. 229, no. 1992, pp. 203–229, 2000.

[13] W. Dai and W. Ji, “A mapreduce implementation of C4. 5 decision tree algorithm,” *Int. J. Database Theory Appl.*, vol. 7, no. 1, pp. 49–60, 2014.

[14] A. An and N. Cercone, “Rule Quality Measures Improve the Accuracy of Rule Induction : An Experimental Approach,” *Found. Intell. Syst. Lect. Notes Comput. Sci.*, pp. 119–129, 2000.

[15] G. P. Zhang, “Neural networks for Classification : a survey,” *IEEE Trans. Syst. Man Cybern. Part C (Applications Rev.*, vol. 30, no. 4, pp. 451–462, 2000.

[16] L. S. Camargo and T. Yoneyama, “Specification of training sets and the number of hidden neurons for multilayer perceptrons,” *Neural Comput.*, vol. 13, no. 12, pp. 2673–80, 2001.

[17] M. a. Kon and L. Plaskota, “Information complexity of neural networks,” *Neural Networks*, vol. 13, no. 3, pp. 365–375, 2000.

[18] C. Neocleous and C. Schizas, “Artificial Neural Network Learning : A Comparative Review,” *Setn 2002, Lnai 2308*, pp. 300–313, 2002.

[19] J. Y. F. Yam and T. W. S. Chow, “Feedforward networks training speed enhancement by optimal initialization of the synaptic coefficients,” *IEEE Trans. Neural Networks*, vol. 12, no. 2, pp. 430–434, 2001.

[20] M. N. H. Siddique and M. O. Tokhi, “Training neural networks: backpropagation vs. genetic algorithms,” *IJCNN01 Int. Jt. Conf. Neural Networks Proc. Cat No01CH37222*, vol. 4, pp. 2673–2678, 2001.

[21] P. Eklund and A. Hoang, “A Performance Survey of Public Domain Machine Learning Algorithms Technical Report,” *Sch. Inf. Technol. Griffith Univ.*, 2002.

[22] Jensen, Finn V, and Thomas Dyhre Nielsen. Bayesian Networks And Decision Graphs. New York: Springer, 2007. Print.

[23] N. Y. Seishi Okamoto, “Effects of domain characteristics on instance-based learning algrithims,” *Theor. Comput Sci*, vol. 298, pp. 207–233, 2003.

[24] J. S. Sánchez, R. Barandela, and F. J. Ferri, “On filtering the training prototypes in nearest neighbour classification,” in *Topics in Artificial Intelligence*, Springer, 2002, pp. 239–248.

[25] K. Veropoulos, C. Campbell, N. Cristianini, and others, “Controlling the sensitivity of support vector machines,” in *Proceedings of the international joint conference on AI*, 1999, pp. 55–60.

[26] S. S. Keerthi and E. G. Gilbert, “Convergence of a generalized SMO algorithm for SVM classifier design,” *Mach. Learn.*, vol. 46, pp. 351–360, 2002.

[27] J. Kivinen, “On the Learnability and Design of Output Codes for Multiclass Problems,” no. 1995, pp. 201–233, 2002.

[28] A. ABBASI and M. YOUNIS, “A survey on clustering algorithms for wireless sensor networks,” *Comput. Commun.*, vol. 30, no. 14–15, pp. 2826–2841, 2007.

[29] C. C. Aggarwal and C. Zhai, *Mining Text Data*. 2012.

[30] J. Brank, M. Grobelnik, and D. Mladenic, “A survey of ontology evaluation techniques,” in *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, 2005, pp. 166–170.

[31] D. Xu, Rui and Wunsch II, “Survey of clustering algorithms,” *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[32] a Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Zomaya, S. Foufou, and A. Bouras, “A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis,” *IEEE Trans. Emerg. Top. Comput.*, vol. 2, no. 3, pp. 1–1, 2014.

[33] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,” *Comput. Geosci.*, vol. 10, no. 2–3, pp. 191–203, 1984.

[34] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: AnEcient DataClusteringMethod forVery Large Databases,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data, Montreal, Canada*, 1996.

[35] A. Hinneburg and D. A. Keim, “An efficient approach to clustering in large multimedia databases with noise,” in *KDD*, 1998, vol. 98, pp. 58–65.

[36] A. Hinneburg and D. A. Keim, “Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering,” 1999.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, vol. 39, no. 1. 1977.

[38] Stiel, Stephanie et al. "Symptoms And Problem Clusters In Cancer And Non-Cancer Patients In Specialized Palliative Care—Is There A Difference?". *Journal of Pain and Symptom Management*48.1 (2014): 26-35.

[39] Masciari, E., G.M. Mazzeo, and C. Zaniolo. "Analysing Microarray Expression Data Through Effective Clustering". Information Sciences 262 (2014): 32-45.

[40] Chen, Chien-Hsing. "A Hybrid Intelligent Model Of Analyzing Clinical Breast Cancer Data Using Clustering Techniques With Feature Selection". Applied Soft Computing 20 (2014): 4-14.

[41] Chao, Cheng-Min et al. "Construction The Model On The Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression And Decision Tree". J Med Syst 38.10 (2014): n. pag.

[42] Mount, David W et al. "Using Logistic Regression To Improve The Prognostic Value Of Microarray Gene Expression Data Sets: Application To Early-Stage Squamous Cell Carcinoma Of The Lung And Triple Negative Breast Carcinoma". BMC Med Genomics 7.1 (2014): 33.

[43] Brims, Fraser J. et al. "A Novel Clinical Prediction Model For Prognosis In Malignant Pleural Mesothelioma Using Decision Tree Analysis". Journal of Thoracic Oncology (2016): n. pag.

[44] Yang, Haofan, and Yi-Ping Phoebe Chen. "Data Mining In Lung Cancer Pathologic Staging Diagnosis: Correlation Between Clinical And Pathology Information". Expert Systems with Applications 42.15-16 (2015): 6168-6176.

[45] Zięba, Maciej et al. "Boosted SVM For Extracting Rules From Imbalanced Data In Application To Prediction Of The Post-Operative Life Expectancy In The Lung Cancer Patients". Applied Soft Computing 14 (2014): 99-108. Web.

[46] Sharaf-El-Deen, Dina A., Ibrahim F. Moawad, and M. E. Khalifa. "A New Hybrid Case-Based Reasoning Approach For Medical Diagnosis Systems". J Med Syst 38.2 (2014): n. pag. Web.

[47] Ping, Xiao-Ou et al. "A Multiple Measurements Case-Based Reasoning Method For Predicting Recurrent Status Of Liver Cancer Patients". Computers in Industry 69 (2015): 12-21. Web.

[48] Mansoori, Eghbal G., and Khadijeh S. Shafiee. "On Fuzzy Feature Selection In Designing Fuzzy Classifiers For High-Dimensional Data". Evolving Systems (2015): n. pag. Web.

[49] Majid, Abdul et al. "Prediction Of Human Breast And Colon Cancers From Imbalanced Data Using Nearest Neighbor And Support Vector Machines". Computer Methods and Programs in Biomedicine 113.3 (2014): 792-808. Web.

[50] Zolbanin, Hamed Majidi, Dursun Delen, and Amir Hassan Zadeh. "Predicting Overall Survivability In Comorbidity Of Cancers: A Data Mining Approach". Decision Support Systems 74 (2015): 150-161. Web.

[51] Kumar, Mukesh, and Santanu Kumar Rath. "Classification Of Microarray Using Mapreduce Based Proximal Support Vector Machine Classifier". Knowledge-Based Systems 89 (2015): 584-602. Web.

[52] Wang, Chang-Yu et al. "Predicting Survival Of Individual Patients With Esophageal Cancer By Adaptive Neuro-Fuzzy Inference System Approach". Applied Soft Computing 35 (2015): 583-590. Web.

[53] Tsai, Meng-Hsiun et al. "A Decision Tree Based Classifier To Analyze Human Ovarian Cancer Cdna Microarray Datasets". J Med Syst 40.1 (2015): n. pag. Web.

[54] Tanha, Jafar, Maarten van Someren, and Hamideh Afsarmanesh. "Semi-Supervised Self-Training For Decision Tree Classifiers". Int. J. Mach. Learn. & Cyber. (2015): n. pag. Web.

[55] NANDHINI, M, and S N SIVANANDAM. "An Improved Predictive Association Rule Based Classifier Using Gain Ratio And T-Test For Health Care Data Diagnosis". Sadhana 40.6 (2015): 1683-1699. Web.

[56] Ghosh, Anupam, and Rajat K. De. "Fuzzy Correlated Association Mining: Selecting Altered Associations Among The Genes, And Some Possible Marker Genes Mediating Certain Cancers". Applied Soft Computing 38 (2016): 587-605. Web.

[57] Kuo, Ren-Jieh et al. "Application Of A Two-Stage Fuzzy Neural Network To A Prostate Cancer Prognosis System". Artificial Intelligence in Medicine 63.2 (2015): 119-133. Web.

[58] Lu, Chunhong, Zhaomin Zhu, and Xiaofeng Gu. "An Intelligent System For Lung Cancer Diagnosis Using A New Genetic Algorithm Based Feature Selection Method". J Med Syst 38.9 (2014): n. pag. Web.

1. Data Mining [↑](#footnote-ref-1)
2. Supervised Learning [↑](#footnote-ref-2)
3. Unsupervised Learning [↑](#footnote-ref-3)
4. Classification [↑](#footnote-ref-4)
5. Support Vector Machine [↑](#footnote-ref-5)
6. Inductive [↑](#footnote-ref-6)
7. Brute-force [↑](#footnote-ref-7)
8. data preparation [↑](#footnote-ref-8)
9. Pre-Processing [↑](#footnote-ref-9)
10. Features [↑](#footnote-ref-10)
11. Don’t Care [↑](#footnote-ref-11)
12. Classifier [↑](#footnote-ref-12)
13. Perceptron-Based Methods [↑](#footnote-ref-13)
14. Bayesian Networks [↑](#footnote-ref-14)
15. Instance-Based Methods [↑](#footnote-ref-15)
16. Logic-Based Algorithms [↑](#footnote-ref-16)
17. Hard [↑](#footnote-ref-17)
18. Disjunctive normal form [↑](#footnote-ref-18)
19. Over Training [↑](#footnote-ref-19)
20. Perceptron-Based Technique [↑](#footnote-ref-20)
21. Neural Network [↑](#footnote-ref-21)
22. Back Propagation [↑](#footnote-ref-22)
23. Feed-Forward Neural Networks [↑](#footnote-ref-23)
24. Instance-Based Learning [↑](#footnote-ref-24)
25. K-Nearest Neighbor [↑](#footnote-ref-25)
26. Support Vector Machine (SVM) [↑](#footnote-ref-26)
27. Quadratic Programming (QP) [↑](#footnote-ref-27)
28. Sequential Minimal Optimization (SMO) [↑](#footnote-ref-28)
29. Naïve Bayes [↑](#footnote-ref-29)
30. Partition-Based [↑](#footnote-ref-30)
31. Hierarchial-Based [↑](#footnote-ref-31)
32. Density-Based [↑](#footnote-ref-32)
33. Statistical INformation Grid [↑](#footnote-ref-33)
34. Fuzzy C-Means [↑](#footnote-ref-34)
35. Balanced Iterative Reducing and Clustering using Hierarchies [↑](#footnote-ref-35)
36. Classification Feature Tree (CF-Tree) [↑](#footnote-ref-36)
37. Clustering Based on Density Distribution Functions "DENsity-based CLUstEring [↑](#footnote-ref-37)
38. Optimal Grid [↑](#footnote-ref-38)
39. Expectation Maximization [↑](#footnote-ref-39)
40. Agglomerative Hierarchical [↑](#footnote-ref-40)
41. Microarray data CLustering Using Binary Splitting [↑](#footnote-ref-41)
42. Filter-Based [↑](#footnote-ref-42)
43. University of California [↑](#footnote-ref-43)
44. Logistic Regression [↑](#footnote-ref-44)
45. Support Vector Machine (SVM) [↑](#footnote-ref-45)
46. Decision Tree [↑](#footnote-ref-46)
47. Receiver operating characteristic [↑](#footnote-ref-47)
48. ANalysis Of VAriance [↑](#footnote-ref-48)
49. National Center for Biotechnical Information Gene Expression Omnibus [↑](#footnote-ref-49)
50. Classification And Regression Tree (CART) [↑](#footnote-ref-50)
51. Sensitivity [↑](#footnote-ref-51)
52. Specificity [↑](#footnote-ref-52)
53. Positive Predictive Value (PPV) [↑](#footnote-ref-53)
54. Association Mining [↑](#footnote-ref-54)
55. Correlation [↑](#footnote-ref-55)
56. Unbalanced Data [↑](#footnote-ref-56)
57. Knowledge Extraction based on Evolutionary Learning [↑](#footnote-ref-57)
58. Case Based Reasoning [↑](#footnote-ref-58)
59. Rule Based Reasoning [↑](#footnote-ref-59)
60. Retrieval Based CBR [↑](#footnote-ref-60)
61. Multi Measure CBR (MMCBR) [↑](#footnote-ref-61)
62. SingleCBR [↑](#footnote-ref-62)
63. Radiofrequency Ablation [↑](#footnote-ref-63)
64. Mega Trend Diffiusion [↑](#footnote-ref-64)
65. K-Nearest Neighbor [↑](#footnote-ref-65)
66. Mega Trend Diffusion – Support Vector Machine [↑](#footnote-ref-66)
67. Mega Trend Diffusion – K-Nearest Neighbor [↑](#footnote-ref-67)
68. Random Forrest [↑](#footnote-ref-68)
69. Artificial Neural Network (ANN) [↑](#footnote-ref-69)
70. Decision Support System [↑](#footnote-ref-70)
71. the Surveillance, Epidemiology, and End Results [↑](#footnote-ref-71)
72. Proximal SVM [↑](#footnote-ref-72)
73. Adaptive Nero Fuzzy Inference System (ANFIS) [↑](#footnote-ref-73)
74. Pathway Analysis [↑](#footnote-ref-74)
75. Classification and Regression Trees [↑](#footnote-ref-75)
76. CHi-squared Automatic Interaction Detection [↑](#footnote-ref-76)
77. Self-Training [↑](#footnote-ref-77)
78. Classification Associative [↑](#footnote-ref-78)
79. Classification Based on Predictive Association Rule [↑](#footnote-ref-79)
80. Fuzzy Neural Network [↑](#footnote-ref-80)
81. An Immune Network Algorithm for Optimisation [↑](#footnote-ref-81)
82. Particle swarm optimization [↑](#footnote-ref-82)
83. Artificial Immune Network + K-means [↑](#footnote-ref-83)
84. Mean Square Error (MSE) [↑](#footnote-ref-84)
85. Genetic Algorithm [↑](#footnote-ref-85)
86. Fitness Function [↑](#footnote-ref-86)
87. Back Propagation Neural Network [↑](#footnote-ref-87)
88. Correlation-Based [↑](#footnote-ref-88)