ISSN: 2745-3673
http://www.Rcsj.ir

مجله علمی پژوهش در علوم رایانه (سال ششم)
شماره ۲۱ / بهار ۱۴۰۰ / ص ۸۹-۶۲

مروری بر داده‌کاوی
(وظایف، تکنیک‌های مورداستفاده و کاربردها)

زهرا نخعی راد

دانشجوی دکتری مدیریت فناوری اطلاعات- کسب و کار هوشمند، دانشکده مدیریت و اقتصاد، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات.

نام نویسنده مسئول:
زهرا نخعی راد

تاریخ دریافت: ۱۴۰۰/۱/۴
تاریخ پذیرش: ۱۴۰۰/۳/۲۷

چکیده

داده‌کاوی که به عنوان علم استخراج الگوها و دانش مفید از پایگاه داده تعریف می‌شود امروزه نقش مهمی در انواع فعالیت‌های انسانی بازی می‌کند؛ و با توجه به قابلیت‌های آن، تبدیل به بخش اساسی در تعداد زیادی از برنامه‌های کاربردی مانند حوزه‌های بانکداری، خرده‌فروشی، پزشکی، بیمه، بیوانفورماتیک و غیره شده است.داده‌کاوی با استفاده از ترکیبی از الگوریتم های یادگیری ماشین، تحلیل‌های آماری، تکنیک‌های مدل‌سازی و تکنولوژی پایگاه داده‌ها، الگوها و روابط دقیق را پیدا می‌کند که پیش بینی آینده را امکان پذیر می‌سازد.به‌منظور اتخاذ یک دیدگاه جامع از روند تحقیق در حوزه داده‌کاوی این مقاله به مروری جامع از کاربردهای داده‌کاوی می‌پردازد و همچنین به بررسی جامع و سیستماتیک از کارها و تکنیک‌های مختلف مورد استفاده در داده‌کاوی از جمله درخت‌های تصمیم، قوانین وابستگی، خوشه‌بندی، سری - های زمانی و چالش‌ها و موضوعات درزمینه تحقیقات داده‌کاوی می‌پردازد.

واژگان کلیدی: داده کاوی، وظایف د اده کاوی، تکنیک های مورد استفاده داده کاوی، کاربردهای داده کاوی

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

داده‌کاوی، ابزاری اساسی و مهم در کشف دانش است و برای استخراج الگوهای ناشناخته مفید موجود در انبار داده‌ها استفاده می‌شود [۳۳]. در واقع داده‌کاوی شامل کارکردها، تکنیک‌ها و الگوریتم‌هایی است که برای کشف و استخراج الگوهای جذاب موجود در داده‌ها استفاده می‌شود [۳۳،۵۱،۷]با توجه به اهمیت در تصمیم‌گیری، در دو دهه اخیر، داده‌کاوی مورد توجه گسترده‌ای قرار گرفته و تبدیل به یک ابزار ضروری در انجام عملیات مختلف سازمان‌ها شده است [۸۲]. داده‌کاوی یک مرحله از کشف دانش در فرآیند پایگاه‌های اطلاعاتی است که شامل کاربرد تجزیه‌وتحلیل داده‌ها و الگوریتم‌های قابل‌قبول است که تحت محدودیت‌های کارایی محاسباتی قابل قبول، تعداد مشخصی از الگوها را از داده‌ها تولید می‌کنند [۳۳،۴۸] داده‌کاوی را "روند کشف یا استخراج الگوهای جالب، انجمن‌ها، تغییرات، ناهنجاری‌ها و ساختارهای مهم از مقدار زیادی داده از چندین منبع جداگانه مانند سیستم فایل، پایگاه داده، انبار داده یا مخازن اطلاعات دیگر " عنوان کردند. داده کاوی یکی از ابزارهای هوشمندی کسب وکار [1] است که اطلاعات و بینش لازم درباره کسب و کار را در اختیار مدیران قرار می‌دهد[۷۶].

بسیاری از تکنیک‌های حوزه‌های دیگر [۴۸] مانند آمار، سیستم‌های پایگاه داده یا انبار داده، یادگیری ماشین، الگوریتم‌ها، تشخیص الگو، تجسم وتصویرسازی، بازیابی اطلاعات، محاسبات با کارایی بالا و غیره نیز در داده کاوی گنجانیده شده است. سه تکنیک اول مشارکت کنندگان اولیه داده‌کاوی است [۳۴]

گرایش در تحقیقات داده‌کاوی

از طریق بررسی ادبیات، مشخص شد که تحقیقات داده کاوی را می‌توان به‌طورکلی به انواع زیر تقسیم‌بندی کرد [۳۷، ۶۴، ۱۰۱، ۱۱۵].

۳.۱ توابع (وظایف) داده کاوی
۳.۲ تکنیک های مورد استفاده در داده کاوی
۳.۳ الگوریتم های مورد استفاده در داده کاوی
۳.۴ دامنه های داده کاوی
۳.۵ برنامه های داده‌کاوی

وظایف داده کاوی

در یک پایگاه داده بزرگ و یا انبار داده ممکن است انواع مختلفی از الگوهای ناشناخته موجود باشد [۱۹] برای استخراج این الگوهای ناشناخته، انواع متمایز وظایف داده‌کاوی و تکنیک‌ها را می‌توان استفاده کرد [۱۰۱، ۱۱۹،۱۱۵]. براساس انواع مختلف الگوها، توابع داده‌کاوی می‌تواند در دسته‌های خلاصه‌سازی، شناسایی و تشخیص، طبقه بندی، رگرسیون و روند تجزیه و تحلیل، خوشه بندی، تجزیه و تحلیل داده‌های پرت و انجمن و غیره می تواند باشد [۳۳،۳۷،۱۷،۳۵]. کار طبقه بندی شده ادبیات فوق الذکر مربوط به وظایف داده کاوی در جدول ۱ ذکر شده است.

[1]. BI(Business Intelligence)

ISSN: 2745-3673
http://www.Rcsj.ir

جدول ۱ دسته بندی مقالات مرتبط با وظایف داده کاوی [۴۶]

وظایف داده کاوی / مقالات	خلاصه سازی	شناسایی و تشخیص	طبقه بندی	خصوصیات انجمنی	خوشه بندی	تحلیل داده های پرت	رگرسیون
Abuaiadah (2015)				✓			
Algergawy et al. (2011)				✓			
Angiulli et al. (2013)						✓	
Angiulli and Fassetti (2016)						✓	
Bhatnagar et al. (2015)		✓		✓		✓	
Bouguessa (2013)				✓		✓	
Caampello et al (2015).				✓		✓	
Carpineto et al. (2009)				✓			
Ceglar and Roddick (2006)					✓		
Chen et al. (1996)	✓	✓	✓	✓	✓		
Chen et al. (2009)					✓		
Chin-Yuan et al. (2012)				✓			
Das et al. (2011)				✓			
Dincer (2006)		✓		✓			
Geng and Hamilton (2006)	✓		✓		✓		
Gupta and Chandra (2019)				✓			
Gupta mg and Chandra (2019)				✓			
HeaZ et al. (2004)			✓			✓	
Hung et al. (2015)					✓		
Hungg and Thu (2016)					✓		
Jain et al. (1999)				✓			
Jin et al. (2014)				✓			

ISSN: 2745-3673
http://www.Rcsj.ir

وظایف داده کاوی / مقالات	خلاصه سازی	شناسایی و تشخیص	طبقه بندی	خصوصیات انجمنی	خوشه بندی	تحلیل داده های پرت	رگرسیون
Khandare and Alvi (2017)				✓			
Koh and Ravana (2016)					✓		
Kosina and Gama (2015) Kotsiantis (2007)			✓				
Kumar et al (2016)			✓				
Lee and Yun (2017)					✓		
Li and Zaki (2015)			✓				
Liao and Triantaphyllou (2007)			✓				
Mabroukeh and Ezeife (2010)					✓		
Mampaey and Vreeken (2011)				✓			
Menardi and Torelli (2012)			✓				
Mukhopadhyay et al. (2015)				✓			
Pei et al . (2016)				✓			
Rafalak et al. (2016)			✓				
Reddy and Jana (2014)				✓			
Rustogi et al (2017)					✓		
Shah-Hosseini (2013)				✓			
Silva et al. (2013)				✓			
Silva and Antunes (2014)					✓		
Sim et al . (2012)				✓			
Sohrabi and Rohani (2017)					✓		
Susan et al.			✓				

ISSN: 2745-3673
http://www.Rcsj.ir

وظایف داده کاوی / مقالات	خلاصه سازی	شناسایی و تشخیص	طبقه بندی	خصوصیات انجمنی	خوشه بندی	تحلیل داده های پرت	رگرسیون
(2006)							
Tan et al. (2009)				✓	✓		✓
Tew et al. (2013)				✓	✓		
Wang and Dong (2015)				✓			
Wang and Sun (2014)			✓	✓			
Wang et al. (2011)				✓			
Zacharis (2018)			✓				
Zhang et al. (2015)				✓			

خلاصه سازی

خلاصه‌سازی به مجموعه کوچکتر منتج می‌شود و خلاصه‌ای از داده‌های دقیق بر اساس مفهوم سلسله مراتب ارائه می دهد. معمولاً خلاصه سازی با استفاده از تجمیع انجام می‌شودکه می تواند به سطوح مختلف انتزاع گسترش یابد واز زوایای مختلف قابل مشاهده است. انواع مختلف الگوها براساس ترکیبات مختلف سطح انتزاع و ابعاد مختلف قابل استخراج است [۱۹]. خلاصه - سازی داده‌ها معمولاً با استفاده از انتساب رویکرد القایی و رویکرد مکعب داده انجام می شود [۴۷،۵۴،۳۹].

رویکرد مکعب داده که به عنوان پایگاه داده چند بعدی نیز شناخته می‌شود اغلب از محاسبات گران قیمت استفاده می‌کند که شامل توابع گروهی است و سپس نتیجه را به صورت دیدگاه های تحقق یافته در پایگاه داده چند بعدی یا ام دی دی بی[۲] جهت پشتیبانی تصمیم و کشف دانش ذخیره می‌کند [۱۹]. رویکرد القایی ویژگی‌گرا داده‌های مرتبط پایگاه داده را با کمک نرم افزارهای اسکیوال [۳] مانند دی ام کیو ال [۴]جمع آوری می‌کند و سپس مجموعه ای از تکنیک ها برای تعمیم داده ها [۴۷] اعمال می شود[۱۹].

شناسایی و تشخیص

شناسایی [۵] اساساً خلاصه سازی داده ها بر اساس سلسله مراتب مفاهیم است. از سویی دیگر، از تشخیص [۶]برای شناسایی انواع مختلف مجموعه داده ها استفاده می شود. خروجی به صورت قوانین تشخیصی تولید می شود [۴۸،۲۶،۱۱].

طبقه‌بندی

طبقه‌بندی فرایند مشاهدات جدید براساس کلاس های از پیش تعیین شده، یعنی یادگیری تحت نظارت است. یک الگوریتم طبقه بندی برای پیش‌بینی کلاس‌هایی از داده استفاده می‌شود [۴۸]. مجموعه بزرگی از الگوریتم های طبقه بندی (یا طبقه‌بندها) تاکنون توسط محققان ارائه شده است [۴۸،۷۲]. برخی از الگوریتم‌های طبقه‌بندی مشهور، در جدول ۲ خلاصه شده است. طبقه بندی بر اساس الگوریتم ژنتیک، مجموعه‌های فازی ویادگیری نیمه نظارت شده فعال توسط برخی از محققان ارائه

[2] Multi-Dimensional Databases (MDDB)
[3] SQL:Structured Query Language
[4] DMQL:Data Mining Query Language
[5] Characterization
[6] discrimination

ISSN: 2745-3673
http://www.Rcsj.ir

شده است [۴۸]. علاوه بر طبقه بندی کننده های معروف ذکر شده درجدول ۲ مجموعه‌ای از طبقه بندهای جدید نیز وجود دارند مانند روش مبتنی بر ویژگی‌های پیش‌بینی با استفاده از یادگیری نظارت شده [۶۷] و طبقه‌بند مبتنی‌بر ویژگی [۲۱] جهت انطباق مقادیر نمادین [۷۸] برای طبقه بندی توزیع انحرافی مشاهدات، قوانین تصمیم گیری بسیار سریع سازگار [7]و غیره [۶۶]

جدول ۲ برخی از الگوریتم های طبقه بندی محبوب [۴۶]

براساس مفهوم	نام الگوریتم ها	دسته بندی
افزایش اطلاعات ضریب شاخص جینی نرخ رشد	ID3 (iterative dichotomiser) based) ID3 (iterative dichotomiser) Information gain CART (classification and regression trees) C4.5 (a descendant of ID3)	طبقه بندی های مبتنی بر تصمیم گیری (مبتنی بر یادگیری ماشین)
طبقه بندی های بیزی (براساس آمار)	Naıve Bayesian (or Simple Bayesian) classifier	تئوری بیزین نظریه بی ز و مدل های گرافیکی احتمال
طبقه بندی های مبتنی بر قانون (مبتنی بر یادگیری ماشین)	IF–THEN rule using decision tree Sequential covering algorithms: AQ, CN2 and RIPPER	درخت تصمیم (بی نظمی) ناهنجاری و افزایش اطلاعات ابر طرح جداکننده بهینه خطی
طبقه بندی ماشین بردار پشتیبان (مبتنی بر یادگیری ماشین)	Support vector machine	شب که ع صبی م صنوعی چن د لای ه FF
طبقه بندی با استفاده از روش پس انتشار (مبتنی بر شبکه عصبی)	Back propagation	استخراج مجموعه داده‌های پر تکرار استخراج و کاوش مجموعه ایتم های پر تکرار با استفاده از استراتژی حرص قاعده
طبقه بندی با استفاده از الگوهای تکرار (مبتنی بر شبکه عصبی)	CBA (classification based on association) CMAR (classification based on predictive association rules)	استخراج و کاوش مجموعه ایتم های پر تکرار به کمک (فویل) foil (first order inductive learner)
یادگیری تنبل (مبتنی بر یادگیری ماشین)	CPAR (classification based on multiple association rules) kNN (K-nearest neighbour) CBR (case-based reasoning)	یادگیری به کمک مقایسه و فاصله اقلیدسی بانک اطلاعاتی از راه حل های مسیله و دانش پس زمینه ای

خوشه بندی یا تجزیه و تحلیل خوشه ای

خوشه‌بندی یا تحلیل خوشه برای تقسیم‌بندی یا بخش‌بندی اشیا داده‌ای یا مشاهدات به زیر مجموعه‌هایی به نام گروه یا خوشه استفاده می‌شود. اشیایی که به هم نزدیکتر و مشابه‌تر هستند در یک گروه قرار می‌گیرند. مانند طبقه بندی، خوشه بندی نیز اشیا داده مشابه را طبقه‌بندی می‌کند اما بر خلاف طبقه بندی در خوشه‌بندی برچسب‌های کلاس ناشناخته هستند به عنوان

[7] adaptive very fast decision rules (AVFDR)

ISSN: 2745-3673
http://www.Rcsj.ir

مثال یادگیری بدون نظارت [۴۸] تجزیه و تحلیل خوشه یکی از محبوب ترین تکنیک ها است که نه تنها در داده کاوی استفاده می شود بلکه در موارد دیگر نیز مانند دامنه هایی مانند آمار، بخش‌بندی تصویر، تشخیص الگو، بازیابی اطلاعات، بیوانفورماتیک و غیره مورد استفاده قرار می‌گیرد [۵۹].

در دو دهه اخیر مجموعه بزرگی از الگوریتم های خوشه بندی توسط بسیاری از محققان پیشنهاد شده است [۳۱،۴۸،۱۰۹]. برخی از الگوریتم‌های خوشه‌بندی محبوب در جدول ۳ ارائه شده اند. الگوریتم‌های خوشه‌بندی بر اساس مدل احتمالات، مجموعه‌های فازی، الگوریتم امید ریاضی بیشینه‌سازی[8]، همبستگی با استفاده از پی سی ای[9] و گراف نیز توسط برخی از محققان پیشنهاد شد [۴۸]. علاوه بر الگوریتم‌های خوشه بندی محبوب ارائه شده در جدول ۳، محققان مجموعه‌ای از الگوریتم‌های خوشه‌بندی جدیدی مانند روش‌های بدون پارامتر با استفاده از حداقل طول توصیف [۷۷] رویکرد خوشه بندی سلسله مراتبی موازی [۱۱۶] روش خوشه بندی داده های بر اساس معیار z-score [۲۴] الگوریتم خوشه بندی کاملاً خودکار داده‌های طبقه بندی شده با ابعاد بالا [۱۳] الگوریتم کامینز مبتنی بر قطره‌های آب هوشمند که ازدحام محور است[10] [۱۰۰] الگوریتم خوشه بندی مبتنی بر نمودار ورنوی برای داده‌های مصنوعی و بیولوژیکی [۹۹]، الگوریتم خوشه‌بندی بایسکت[11] [۱] الگوریتم خوشه‌بندی چگالی محور مبتنی بر دانش[12] [۶۰] الگوریتمی برای خوشه‌بندی مجموعه داده مقیاس بزرگ بر اساس ترکیب منحصر به فرد تجزیه ماتریس وتقریب ماتریس درجه پایین[13]به نام تجزیه ماتریس پراکنده سطح پایین[14] [۱۱۸] روش خوشه‌بندی سه مرحله‌ای بر اساس آنالیز افتراقی[15] [۱۱] وغیره را پیشنهاد کرده اند. کمپلو و همکاران نیز در مقاله [۱۴] چارچوبی برای خوشه‌بندی مبتنی بر چگالی ارائه کردند. خانداره و الوی نیز الگوریتم خوشه بندی بهبود یافته ای را با ارائه روش جدیدی جهت بهبود مقداردهی اولیه خوشه‌ها پیشنهاد دادند [۶۳].

جدول ۳ برخی از الگوریتم های خوشه بندی محبوب [۴۶]

براساس مفهوم	نام الگوریتم ها	دسته بندی
روش افتراقی روش متراکم مدل پویا خوشه بندی درخت ویژگی مدل احتمالی	DIANA (divisive analysis) AGNES (agglomerative nesting) Chameleon BIRCH (balanced iterative reducing and clustering using hierarchies) Probabilistic hierarchical clustering	سلسله مراتبی
مرکز شی نماینده نمونه برداری نمونه برداری تصادفی شی نماینده	k-Means k-Medoids CLARA (clustering large applications) CLARANS (clustering large applications based upon randomized search) PAM (partitioning around medoids)	مبتنی بر پارتیشن بندی
شناسایی یکنواخت سلول های متراکم با توجه به ابعاد	CLIQUE (clustering in quest) STING (statistical information grid)	مبتنی برش بکه

[8] expectation–maximization
[9] PCA:PRINCIPAL COMPONET ANALYSIS
[10] IWD-KM: Intelligent Water Drops K-means
[11] bisect K-means clustering algorithm
[12] domain knowledge based density-based clustering
[13] combination of matrix decomposition and low-rank matrix approximation
[14] exemplar-based low-rank sparse matrix decomposition (EMD)
[15] روش‌های آماری هستند که از جمله در یادگیری ماشین و بازشناخت الگو برای پیدا کردن ترکیب خطی خصوصیاتی که به بهترین صورت دو یا چند کلاس از اشیا را از هم جدا می‌کند، استفاده می‌شوند.

ISSN: 2745-3673
http://www.Rcsj.ir

سلول های شبکه ایحاوی اطلاعات اماری		
مناطق متصل با تراکم بالا مناطق متصل با تراکم زیاد مشخص شده به وسیله پارامتر های چگالی سراسری تابع توزیع احتمالی	DBSCAN (density-based spatial clustering of application of noise) OPTICS (ordering points to identify the clustering structure) DENCLUE (density-based clustering)	مبتنی بر تراکم

گوپا و چاندرا یک رویکرد کارآمد مبتنی بر انتخاب کاملاً تفکیک شده نقاط داده به عنوان مرکز خوشه اولیه جهت بهبود عملکرد الگوریتم کامینز ارائه کردند.روش‌های جدید مقداردهی اولیه خوشه‌ها در کامینز که با استفاده از پارتیشن‌بندی انجام می‌شود را به ترتیب پی-کامینز [16] و ام پی-کامینز [17] می‌نامند که به ترتیب در مراجع [۴۲] توسط گوپا و چاندرا ارائه گردید. مقداردهی اولیه خوشه بر اساس روش مکعب که توسط گوپا و چاندرا پیشنهاد شد، هیبسیم[18]نامیده می شود [۴۲]. الگوریتم‌های هیبسیم، پی-کامینز و ام پی-کامینز در مقایسه با روش‌های سنتی نتایج بهتری می دهند.

خوشه بندی داده های ایکس ام ال[19]یکی از مشکلات ساده در بسیاری از برنامه های داده کاوی مانند وب کاوی، پردازش کوئری‌های ایکس ام ال، بیوانفورماتیک و غیره است. روش های متداول و مرسوم خوشه بندی‌داده‌ها برای خوشه‌بندی داده‌های ایکس ام ال مناسب نیستند [۲].

تکنیک‌های خوشه‌بندی سنتی برای خوشه‌بندی نتایج جستجوی وب مناسب نیستند زیرا نیازهای خاصی دارند [۱۵]. خوشه‌بندی جریان داده‌ها نیز فرایند دشواری است و نیاز به توانایی خوشه‌بندی پیوسته و مداوم اشیاء جریان‌دار در محدودیت-های حافظه و زمان داده شده دارد [۱۰۲].

تجزیه و تحلیل داده‌های پرت

اشیا داده ای که در رفتار کلی داده‌ها متفاوت هستند، به عنوان داده‌های پرت [20] خوانده می‌شوند. در کل داده‌های پرت توسط بیشتر روش‌های داده‌کاوی به عنوان نویز یا استثنا کنار گذاشته می شوند.گاهی اوقات، ممکن است داده‌های پرت اطلاعات بیشتری در مقایسه به سایر اشیا داده‌ای داشته باشند بنابراین تجزیه و تحلیل داده‌های پرت در برخی از برنامه‌های کاربردی مانند کشف نفوذ، کشف تقلب، تشخیص ناهنجاری و غیره مهم هستند [۴۹].

بسیاری از تکنیک‌های داده‌کاوی معمولاً از خوشه‌بندی برای تشخیص داده‌های پرت به عنوان نویز استفاده می‌کنند. روش های تشخیص داده‌های پرت می‌توانند به دسته‌های روش های مبتنی بر طبقه‌بندی، روش‌های آماری، روش‌های مبتنی بر خوشه بندی، روش‌های نظارت‌شده، نیمه نظارت شده، روشهای بدون نظارت، روش‌های مبتنی بر انحراف و روشهای مبتنی بر فاصله طبقه بندی کرد [۴۸].

انجولی و فستی اظهار داشتند که دانش پس‌زمینه یا دانش دامنه‌ای می‌تواند به راحتی جهت شناسایی داده‌های پرت استفاده شود.آنها راه حلی بدون نظارت پیشنهاد کردند که می‌تواند با یادگیری نظارت شده رابطه داشته باشد. عامل داده پرت گرادیان ارائه شده توسط انجولی و فستی برای تعمیم و ادغام داده‌های پرت آماری بررسی شده است [۴]. کامپو و همکاران چارچوبی مبتنی بر چگالی برای تشخیص داده‌های پرت ارائه دادند[۱۴].دو الگوریتم جدید inc-iVAT و dec-iVAT بر اساس ارزیابی بصری تمایل به تشخیص ناهنجاری در جریان داده ها معرفی شده‌اند [۶۸].

---

16 P-K-Means
17 M-P-K-Means
18 Hybcim:Hypercube Based Cluster Initialization Metod
19 Xml:Extensibel Markup Language
20 Outlier

ISSN: 2745-3673
http://www.Rcsj.ir

تجزیه و تحلیل انجمن‌ها یا استخراج انجمن‌ها

تجزیه و تحلیل انجمنی ارتباطات و پیوندها را در میان مجموعه داده ها و اشیا داده‌ای که می‌توانند به صورت انجمنی حداقل آستانه پشتیبانی و حداقل آستانه اطمینان را ارضا نماید را شناسایی می‌کند. شناسایی مجموعه اقلام تکراری با تولید قوانین انجمنی قوی را کاوش قوانین انجمنی گویند [۴۸،۲۰].

تجزیه و تحلیل انجمن‌ها شامل استخراج مجموعه داده‌های تکراری، زیردنباله‌ها و زیرساخت‌ها است [۴۸]. تجزیه و تحلیل سبد خرید به طور عمده است با استفاده از تجزیه و تحلیل انجمن ها انجام می‌شود. الگوریتم اپریوری[۲۱] به طور گسترده ای است برای انجمن‌ها استفاده می‌شود. الگوریتم های تجزیه و تحلیل انجمن را می‌توان در الگوریتم‌های کلاسیک، الگوریتم‌های نمایش متراکم و الگوریتم‌های مجموعه‌های ناقص طبقه‌بندی کرد [۱۶].

برخی از الگوریتم های معروف کاوش انجمنی به طور خلاصه در جدول ۴ بررسی شده است. الگوریتم‌های کاوش انجمنی برای انجمن‌های چندسطحی، انجمن‌های چندبعدی، انجمن‌های کمی، الگوهای نادر، انجمن‌های مبتنی بر محدودیت و غیره نیز بوسیله بعضی محققان پیشنهاد شده است [۴۸].

داده کاوی چند رابطه‌ای[۲۲] فرایندی است که برای جستجو الگوهای مبتنی بر جدول چندگانه از آن استفاده می‌شود [۱۰۳]. روشهای نمونه‌گیری با استفاده از فرم نرمال انحصاری[۲۳] توسط لی و زاکی توسعه یافته است [۷۴]. امروزه کاوش الگوهای کم تکرار و نادر در برخی از برنامه ها محبوب شده است [۶۵]. یک الگوریتم صحیح و کارآمد برای کاوش الگوهای پرتکرار با استفاده از حداقل ساختار داده توسط لی و یون بررسی شد [۷۰]. یک الگوریتم کم زمانبرتر مبتنی بر اتوماتای یادگیر سلولی[۲۴] برای استخراج اقلام پرتکرار توسط سهرابی و روشنی ارائه شده است [۱۰۷].

داده کاوی همچنین می‌تواند الگوها و قوانین غیرجالب را نیز استخراج کند. بنابراین ارزیابی الگو (اندازه گیری جذابیت) فقط برای فیلترکردن الگوهای جالب مورد نیاز است. جن و هامیلتون نه مورد خاص را جهت اندازه‌گیری جالب بودن قوانین استخراج شده و خلاصه ارائه کردند بعلاوه این نه معیار به سه دسته (۱) ذهنی، (۲) عینی و (۳) معنایی محور طبقه بندی شده - اند. تیو و همکاران تکنیکی برای تعادل بین معیارهای جذابیت با استفاده از خوشه بندی مبتنی بر رفتار رتبه بندی قانون پیشنهاد کردند [۱۱۰]. هانگ و همکارانش روش FPGrowth مبتنی بر جریان WSWFP [۲۵] را برای استخراج آیتم‌های پرتکرار وزن‌دار برای جریان داده ها را پیشنهاد کردند. یک الگوریتم جدید به نام MFIWDSIM [۲۶] مبتنی بر وزن با استفاده از ماتریس معکوس برای استخراج موارد تکراری ارائه شده است [۵۶]. روستاگی و همکاران الگوریتم اپریوری موازی بهبود یافته را برای چند هسته‌ای ارائه دادند[۹۶].

جدول ۴ برخی از الگوریتم های رایج انجمنی [۴۶]

براساس مفهوم	نام الگوریتم ها	دسته بندی
نسل کاندید مفید	Apriori	اپریوری مانند
مبتنی بر الگوی شرطی بدون نسل کاندید	FP-growth	مبتنی بر رشد الگوهای تکراری
انتقال داده و نسل کاندید	Eclat (equivalence class transformation)	مبتنی بر قالب دادهای عمودی

[21] Apriori
[22] MRDM: Multi-relational Data Mining
[23] DNF: disjunctive normal form
[24] CLA: cellular learning automata
[25] WSWFP
[26] MFIWDSIM: mining frequent ITEMSETS with weights for data stream

ISSN: 2745-3673
http://www.Rcsj.ir

رگرسیون و تحلیل روندیا تکامل

تحلیل و بررسی رگرسیون با گذشت زمان مقدار ویژگی را بر اساس تکنیک‌های رگرسیون پیش بینی می‌کند. مقادیر آینده متغیرها با کمک سری‌های زمانی تاریخچه گذشته پیش‌بینی می‌شوند [۴۸]. تجزیه و تحلیل روندکه به آن تجزیه و تحلیل تکاملی نیز گفته می شود الگوهای جالب را در تاریخچه تکامل اشیاء کشف می‌کند. شناسایی الگوها در تکامل یک شی و مطابقت روند تغییر اشیا دو جنبه اصلی تحلیل روند است [۲۰]. روند اشیا، که رفتار آنها با گذشت زمان تکامل می یابد، می تواند با استفاده از تجزیه و تحلیل روند و مدل های رگرسیون توصیف شود. تحلیل روند روندهای متغیر با زمان را از اشیا داده داخل مجموعه داده نشان می‌دهد. همچنین می‌توان از تجزیه و تحلیل انجمن برای تحلیل تکامل استفاده کرد [۱۰۹].

تکنیک های مورد استفاده در داده کاوی

از آنجا که داده کاوی یک زمینه چند رشته ای است، تکنیک‌ها و روش‌های گوناگونی از تعدادی دامنه در داده‌کاوی اتخاذ می‌شود که شامل آمار، یادگیری ماشین، شبکه های عصبی، سیستم‌های پایگاه داده، الگوریتم‌های ژنتیک، مجموعه‌های فازی، بصری سازی و غیره است. [۳۱،۳۷،۴۸]. ادبیات طبقه‌بندی شده مربوط به تکنیک‌های داده‌کاوی در جدول ۵ ذکر شده است.

جدول ۴ دسته بندی مقالات مربوط به تکنیک های مورد استفاده داده کاوی [۴۶]

تک نیک ه ای مورد استفاده در داده کاوی	آم ار	یادگیری ماشین	شبکه عصبی	پایگاه دادها، انبار داده	الگوریتم ژنتیک	مجموعه منطق فازی	تجسم و بصری‌سازی
Abuaiadah(2015)	✓						
Angiulli et al. (2013)		✓	✓				
Angiulli and Fassetti(2016)	✓						
Bhatnagar et al.(2015)	✓						
Bouguessa(2013)	✓						
Caampello et al(2015).	✓						✓
Carpineto et al. (2009)				✓			
Chen et al.(1996)				✓			
Chen et al. (2009)	✓						
Chin-Yuan et al. (2012)		✓	✓				
David et al. (2005)		✓					
Das et al. (2011)	✓						
Edward and Olgierd (2011)		✓					
Esling and Agon (2012)	✓			✓			
Eyke		✓				✓	

ISSN: 2745-3673
http://www.Rcsj.ir

تک نیک ه ای مورد استفاده در داده کاوی	آم ار	یادگیری ماشین	شبکه عصبی	پایگاه دادها، انبار داده	الگوریتم ژنتیک	مجموعه منطق فازی	تجسم و بصری‌سازی
(2005)							
Eyke (2011)		✓				✓	
Friedman (1997)	✓					✓	
Gengshen and Guenther (2014)		✓					
Jain et al (1999)		✓	✓		✓	✓	
Jin et al. (2014)	✓						
Kate et al. (2000)			✓				
Koh and Ravana (2016)	✓	✓	✓				
Kosina and Gama (2015)		✓					
Kotsiantis (2007)		✓	✓				
Kumar et al. (2016)		✓					
Lee and Yun (2017)		✓					✓
Mabroukeh and Ezeife(2010)				✓			
Mampaey and Vreeken(2011)	✓			✓			
Menardi and Torelli(2012)	✓	✓					
Mukhopadhyay et al.(2015)					✓		
Mu-Jung et al. (2006)			✓			✓	
Padhraic (2000)	✓						
Pei et al . (2016)		✓					
Philip and Salvatore (1997)		✓					
Rafalak et al.(2016)	✓						
Saeed and			✓				

ISSN: 2745-3673
http://www.Rcsj.ir

تجسم و بصری‌سازی	مجموعه منطق فازی	الگوریتم ژنتیک	پایگاه داده‌ها، انبار داده	شبکه عصبی	یادگیری ماشین	آم ار	تک نیک ه ای مورد استفاده در داده کاوی
							Ali(2012)
			✓				Silva et al. (2013)
						✓	Sim et al .(2012)
					✓		Singh et al(2007)
					✓		Sohrabi and Rohani(2017)
					✓		Susan et al. (2006)
		✓					Tan et al.(2009)
						✓	Tew et al.(2013)
					✓	✓	Wang and Dong(2015)
				✓		✓	Wang and Sun (2014)
	✓				✓		Zhang et al. (2015)

در ادامه جدول ۶ مقایسه تمام تکنیک‌های مورد استفاده در داده‌کاوی را نشان می‌دهد [۶].

جدول ۵ مقایسه بین تکنیک های مورد استفاده در داده کاوی [۶]

تکنیک مورد استفاده در داده‌کاوی	ویژگی‌ها	مزایا	محدودیت‌ها
یادگیری قانون وابستگی[۲۷]	• الگوهای مشابه را پیدا می‌کند و قوانینی تولید می‌نماید. • از داده‌ها ارتباطات وابستگی تولید می‌کند. • به فرایند تصمیم‌گیری کمک می‌کند. • از حداقل مقدار پشتیبان و اطمینان استفاده می‌کند.	• به یافتن الگوهای دنباله‌ای کمک می‌کند. • از روش‌های بررسی یکپارچگی، ادغام و کسب استفاده می‌کند.	• الگوهای منطقی با متغیر وابسته نشان نمی‌دهد. • ارزش پشتیبان و اطمینان پیش نیاز هستند.
طبقه‌بندی(کلاس- بندی)	• تکنیکی که قوانینی را پیدا می‌کند که داده‌ها را به گروه‌های مختلف تقسیم می‌کند. • مشاهدات مشابه از پایگاه داده عظیم را شناسایی کرده و آنها را در یک مجموعه مرتب می‌کند	• کارآمدی خوبی دارد و داده پرت را دسته می‌کند. • برای طبقه‌های چند مدلی مناسب است و دفعات محاسبتی کوتاهی نیاز دارد.	• به ساختار محلی داده‌ها حساس است و حافظه زیادی نیاز دارد.
تحلیل خوشه	• تکنیکی برای گردآوری موارد در یک	• نمای سطح بالایی از آنچه در	• به محض انجام ادغام یا

[27]. association rule learning

ISSN: 2745-3673
http://www.Rcsj.ir

تکنیک مورد استفاده در داده‌کاوی	ویژگی‌ها	مزایا	محدودیت‌ها
	گروه است و ویژگی‌های مشابه را پیدا می‌کند. • هدف این تکنیک بدین قرار است: ✓ کشف گروهبندی طبیعی ✓ تولید فرضیه از داده‌ها برای یافتن سازماندهی‌های موثق از داده‌ها	پایگاه داده در جریان است برای کاربر فراهم می‌کند. • تکنیک بسیار کارآمدی است.	جداسازی، قابل تصحیح یا برگردانی نیست.
درخت تصمیم	• مدلی که برای پیش‌بینی کاربرد دارد و می‌تواند به شکل یک درخت در نظر گرفته شود. • ساختاری مانند نمودار جریان(فلوچارت) دارد. • هر شاخه از درخت بیانگر شرط است و برگ‌ها نتایج را اگر شرط احراز شود، بازنمایی می‌کند. • درخت تصمیم داده‌ها را با توجه به شرط تقسیم‌بندی می‌کند و به تصمیم‌گیری کمک می‌نماید. درخت تصمیم می‌تواند بعنوان ابزار کمکی تصمیم بکار رود.	• فهم و تفسیر آن آسان است. • قادر به دسته‌بندی داده های عددی و مقوله‌ای است. • قدرتمند است • در دیتاست‌های عظیم عملکرد خوبی دارد.	• گاهی محاسبات پیچیده است. • گاهی مشکل برازش بیش از حد دارد.
شبکه عصبی	• برای تشخیص الگوهای متفاوت کاربرد دارد. بروندادهای عددی تولید می‌کند. • در شناسایی کلاهبرداریها، عکس - العمل مشتری، درک تصویری و غیره کاربرد گسترده‌ای دارد.	• قابلیت یادگیری خوبی دارد. • سرعت خوبی دارد.	• تنها داده‌های عددی را دسته می‌کند. بنابراین باید هر داده را به شکل عددی ترجمه کنیم. • به خاطر یادگیری ممکن است مشکل بهینه سازی محلی پیش بیاید.
ماشین بردار پشتیبان[۲۸]	• تکنیک یادگیری با نظارت • به حداقل سازی ریسک و حداقل سازی خطای طبقه‌بندی کمک می‌کند. • بخش از کلاس بندی خطی است و میتواند امتدادی از تکنیک پیش‌بین باشد.	• طبقه بندهای بسیار صحیحی تولید می‌کند. • برازش بیش از حد کمی دارد و داده پرت را بکار می‌برد. • حافظه فشرده‌ای دارد و در شناسایی کاراکترهای دست نوشته مفید است.	• سرعت پایینی در آموزش و آزمون دارد شاید به دلیل وجود داده‌های مجزاست. • پیچیدگی الگوریتمی بالایی دارد.
تحلیل رگرسیون	• ارتباط بین دو متغیر متفاوت را تحلیل می‌کند • درمی‌یابد که چگونه ارزش یک متغیر وابسته با تغییر ارزش متغیر مستقل تغییر می‌کند.	• به تصحیح خطاها کمک می‌کند • به پشتیبان‌های پیش‌بین در تصمیم‌گیری کمک می‌کند. • بر رابطه بین متغیر وابسته و مستقل دلالت دارد.	• عدم ثبات پارامترها • توزیع عمومی روابط

[28] Support vector machines

ISSN: 2745-3673
http://www.Rcsj.ir

تکنیک مورد استفاده در داده‌کاوی	ویژگی‌ها	مزایا	محدودیت‌ها
	• از تابعی برای متغیر مستقل استفاده میکند که به تابع رگرسیون معروف است.		
کاوش قواعدانجمنی	• الگوهای گوناگون را از پایگاه داده پیدا می‌کند. • روندهای جاری و یا وقوع قاعده‌مند حوادث مشابه را پیدا می‌کند. از الگوریتم اپریوری استفاده می‌کند.	• به پیش‌بینی براساس حوادث متوالی کمک می‌کند. • مکانیزم یادگیری ساده‌ای دارد.	• گاهی به زمان محاسباتی زیادی نیاز دارد. • ظرفیت پیش‌بینی محدودی دارد.

رویکردهای آماری

گاهی اوقات اصطلاحات "آمار" یا "تکنیک های آماری" به عنوان نام مستعار برای داده کاوی استفاده می شود. اما، آمار قبل از اصطلاح "داده کاوی" ساخته شد آمار داده محور است و برای کشف الگوها و ساخت مدل پیش‌بینی استفاده می‌شود (در آمار که به عنوان رگرسیون نیز نامیده می‌شوند). با توجه به روش‌های مشتق شده از داده، آمار نیز به عنوان یکی ازعمده ترین تکنیک داده کاوی استفاده می‌شود [۸۱]. به عبارت دیگر، داده‌کاوی با آمار ارتباط ذاتی دارد [۴۸]. بسیاری از ابزار تجزیه و تحلیل آماری از جمله شبکه بیزی، همبستگی تجزیه و تحلیل، تحلیل عامل، تحلیل تفکیک، تجزیه و تحلیل خوشه‌ای، تحلیل رگرسیون و غیره به طور گسترده ای برای داده کاوی استفاده می شود [۳۴،۳۵،۵۷]. معمولاً بیشتر مدل های آماری از مجموعه داده‌های آموزشی ساخته می شوند. الگوها و قوانین گوناگون از روی مدل ترسیم می‌شود. بیشتروظایف داده کاوی با استفاده از یک یا چند رویکرد آماری انجام می شود [۳۵].

روش های آماری که معمولاً در داده کاوی استفاده می شود به صورت زیر شرح داده شده است [۳۴، ۳۵، ۵۷].

- شبکه بیزی: نشان دهنده رابطه سببی در میان متغیرها، است، از طریق تئوری بیزین محاسبه شده است [۵۰].
- همبستگی: رابطه بین ابعاد دو یا چند متغیر، واقعیت یا بعد را می توان با استفاده از همبستگی تعیین کرد [۹۱].
- رگرسیون : مشتق یک تابع برای نگاشت مجموعه‌ای از متغیرهای اشیا مختلف به یک متغیر خروجی است [۱۲۰]
- تجزیه و تحلیل خوشه ای: اشیا را بر اساس معیار شباهت گروه بندی می‌کند به طوری که اشیایی که مشابه هستند در یک خوشه قرار می‌گیرند [۳].
- آنالیز افتراقی: اشیا داده را به یک یا گروههای بیشتری بر اساس تابع افتراق نسبت می‌دهد [۸۰].
- تحلیل عاملی: جهت شناسایی و درک دلایل اصلی برای همبستگی و تعریف مهم ترین آنها استفاده می‌شود [۱۱۱].

یادگیری ماشین

یادگیری ماشین مشخص می‌کند که چگونه ماشین‌ها و انسان‌ها می‌توانند از داده‌ها بیاموزند. با توجه به اهمیت یادگیری ماشین در داده کاوی، تعداد زیادی از الگوریتم‌های داده کاوی ریشه در یادگیری ماشین دارند [۸۸].

یادگیری ماشین باعث افزایش سطح خودکار در کشف دانش مربوط به فرآیند پایگاه داده برای بهبود دقت و کارایی می-شود. سیستم های تولید شده توسط یادگیری ماشین می تواند به طور منظم در بخش صنعت یا آموزش استفاده شود در برخی از کاربردها، یادگیری ماشین عملکرد بهتری نسبت به روش های بدون یادگیری دارند [۶۹، ۱۰۴].

استقرایی و قیاسی دو دسته از یادگیری ماشین هستند. یادگیری قیاسی به حقایق و دانش موجود در طول زمان می‌پردازد و سپس دانش جدیدی از دانش قدیمی ایجاد می‌کند. در یادگیری استقرایی، مثالها تعمیم داده می شوند. فرا یادگیری تعدادی از فرایندهای یادگیری مجزا را در یک مد ذهنی ترکیب می‌کند [۱۸].

ISSN: 2745-3673
http://www.Rcsj.ir

یک معماری فرا یادگیری دو رفتار اصلی را نشان می دهد: (۱) یک سیستم دقیق طبقه بندی نهایی (یا نتیجه نهایی) (۲)و باید، نسبت به یک الگوریتم یادگیری ترتیبی منحصربفرد سریع باشد [۱۸].

برای شناسایی و استخراج راه‌حل‌های DSS [۲۹] الگوریتم RSA [۳۰](تجزیه و تحلیل سلسله مراتبی) و DNA [۳۱] (تجزیه و تحلیل شبکه وابستگی[۳۲]) توسط گنشن و گیونتر پیشنهاد شده‌است [۲۷].

افزایش محبوبیت اینترنت منجر به افزایش حملات شبکه شده است. بنابراین، تشخیص نفوذ[۳۳] در حال تبدیل شدن به یکی از مهمترین حوزه‌های تحقیقاتی برای امنیت شبکه می‌شود این روش برای شناسایی دسترسی‌های غیر مجاز استفاده می شود. یادگیری ماشین در سیستم های تشخیص نفوذ استفاده می‌شود. سیستم‌های تشخیص نفوذ امنیت رایانه‌ها را کنترل کرده و هشدار برای گزارش هرگونه تخلف فعال می‌کند [۱۱۳]. این هشدارهای گزارش شده برای ارزیابی و اقدام مناسب به تحلیلگر داده تحویل داده می‌شوند. چی فانگ و همکاران دو روش مبتنی بر کاهش تعداد مثبت کاذب در تشخیص نفوذ ارائه نمودند [۱۱۳].

شبکه عصبی

شبکه عصبی یک شبکه یا مدار شناختی از نورون‌ها است. شبکه عصبی این قابلیت را دارد که با مثال یاد بگیرد و همین آنها را انعطاف پذیر و قدرتمند می‌کند. شبکه‌های عصبی مصنوعی[۳۴] از نورون‌های مصنوعی یا گره‌ها و سیگنال‌های الکتریکی مشابه شبکه های عصبی بیولوژیکی تشکیل شده است [۷۳]. در شبکه‌های عصبی مصنوعی، دانش به صورت لایه‌ای نشان داده می‌شود، مجموعه ای از پردازنده‌های بهم پیوسته که به آنها نورون نیز گفته می‌شود. انواع مختلفی از مدل‌های شبکه عصبی جهت حل مشکلات استفاده می‌شود و به عنوان ابزار تحقیق عملیاتی مدرن نقش حیاتی ایفا می کند [۱۰۶].

طبقه‌بندی بر اساس شبکه‌های عصبی مصنوعی برای پیش بینی موثر مقادیر در آینده توسط دیوید و همکاران بررسی شد [۲۹]. سعید و علی یک پروتکل جدید حفظ حریم خصوصی برای داده های پارتیشن بندی شده بر اساس یادگیری ماشین شدید[۳۵]و پس انتشار[۳۶] ارائه دادند [۹۸] از شبکه‌های عصبی مصنوعی می‌توان در تجزیه و تحلیل داده‌های محیطی نیز استفاده کرد. کانوسکی و همکاران ارزشیابی و نمایش داده‌ها را بررسی کردند [۶۱]. مدل ترکیبی مبتنی بر رگرسیون بردار پشتیبان و الگوریتم‌های یادگیری ماشین پرسپترون چند لایه[۳۷] به عنوان شبیه سازی های متوالی یادگیری ماشین[۳۸]نیز توسط کانوسکی و همکاران پیشنهاد گردید [۶۱].

سیستم های پایگاه داده و انبارهای داده

اگرچه رویکردهای پایگاه‌داده گرا و انبارداده محور بر اساس بهترین مدل‌ها نیستند اما از مدل‌های داده جهت بهره‌برداری از مشخصات داده‌ای موجود استفاده می‌کنند [۱۹]. جهت دستیابی به مقیاس‌پذیری و اثربخشی بیشتر وظایف داده کاوی که نیاز به مدیریت مجموعه داده های بزرگ دارند می‌توان از فناوری‌های داده کاوی استفاده کرد. قابلیت‌های تحلیل داده سیستماتیک در سیستم‌های پایگاه داده تجاری نیز تعبیه شده است[۳۳]. اسکن‌های تکراری بانک اطلاعاتی برای تمرکز بر ویژگی، استنتاج ویژگی محور ومجموعه اقلام پرتکرار عمده ترین روشها برای این رویکرد هستند[۲۰]. ماهیت چند بعدی داده ها ساختار در انبار داده ها نیز داده‌کاوی چند بعدی را ارتقا می دهد [۴۸].

---

[29] DSS: A decision support system
[30] RSA: rough set analysis
[31] dependency network analysis
[32] DNA: dependency network analysis
[33] IDS: intrusion detection system
[34] ANN: Artificial neural networks
[35] ELM: extreme learning machine
[36] BP: back-propagation
[37] ML: Meta Language
[38] MLRSS: MACHIN learning residuals sequential simulations

ISSN: 2745-3673
http://www.Rcsj.ir

الگوریتم ژنتیک

الگوریتم‌های ژنتیک و فرایندهای انتخاب، تولید مثل، جهش و بقا بر اساس مفهوم ارزیابی بیولوژیکی طبیعی ساخته شده است. الگوریتم‌های ژنتیک درست مثل طبیعت با ترکیب دی ان ای موجودات زنده می‌توانند راه حل بهتری ارائه دهند [۵۸]. اما، در ژنتیک توضیح الگوریتم‌ها دشوار است و اندازه گیری آماری وجود دارد تا کاربر بتواند درک کند چرا به راه حل خاص رسیده است [۵۷].

مجموعه های فازی

مفهوم تئوری مجموعه‌های فازی توسط لطفی زاده بنیان گذاری شد. مجموعه فازی درجه عضویت را بر اساس مقدار احتمال محاسبه شده با کمک تابع عضویت تعریف می‌کند و به صورت گسترده‌ای در طبقه بندی وتجزیه و تحلیل خوشه ای استفاده می‌شود [۴۸]. نظریه مجموعه‌های فازی در حال ایجاد پتانسیل مشارکت در برنامه‌های مختلف داده‌کاوی، یادگیری ماشین و زمینه های مرتبط است [۵۳،۵۲]. یک مدل کشف دانش مبتنی بر ادغام اصلاحات تراکنش‌های فازی الگوریتم داده کاوی[۳۹]و استنتاج فازی مبتنی بر شبکه سازگار[40] توسط موجانگ و همکاران توصیف شده است [۵۵]. یک رویکرد یادگیری ماشین ترکیب شده با مدل سازی فازی که مجموعه ای از قوانین فازی را برمی گرداند توسط ادوارد و همکاران پیشنهاد شده است [۷۹].

تجسم و بصری‌سازی

تجسم و بصری‌سازی یک روش داده کاوی بسیار مفید جهت تعریف و نمایش الگوها در مجموعه داده‌ها است. در تجسم و بصری‌سازی داده‌ها در فضای دو یا سه بعدی به اشیایی مانند نقاط، خطوط، مناطق و غیره تبدیل می‌شوند. با بررسی بصری-سازی، کاربران می‌توانند الگوهای جالبی را به طور تعاملی استخراج کنند [۳۵،۸۱]. کامپلو و همکاران چارچوبی برای تخمین بصری‌سازی مبتنی بر تراکم ارائه دادند[۱۴].

رابطه بین وظایف داده کاوی و تکنیک‌های مورد استفاده داده‌کاوی

وظایف داده‌کاوی با یک یا چند تکنیک انجام می‌شود. در عملیات داده درکاوی، یک یا چند تکنیک استفاده می‌شود. جدول ۷ نشان می‌دهد عملیات داده‌کاوی بر اساس کدام تکنیک‌های مورد استفاده داده‌کاوی قابل انجام است.

جدول ۷ ارتباط بین وظایف داده کاوی و تکنیک های مورد استفاده در داده کاوی [۸]

تکنیک های داده کاوی	خلاصه سازی	شناسایی و تشخیص	طبقه بندی	خوشه بندی	خصوصیات انجمنی	تحلیل داده های پرت	رگرسیون و تحلیل ترند
آمار	✓	✓	✓	✓	✓	✓	✓
یادگیری ماشین		✓	✓	✓	✓	✓	✓
شبکه عصبی		✓	✓	✓	✓	✓	✓
سیستم بانک اطلاعاتی	✓	✓			✓	✓	
الگوریتم ژنتیک			✓	✓	✓	✓	
مجموعه منطق		✓	✓	✓	✓	✓	

[39] MFTDA: modification of the fuzzy transaction data-mining algorithm
[40] ANFIS: adaptive-network-based fuzzy inference system

ISSN: 2745-3673
http://www.Rcsj.ir

فازی							
بصری سازی		✓	✓	✓		✓	✓

برنامه‌های کاربردی داده‌کاوی در دنیای واقعی

با توجه به قدرت داده‌کاوی برای تجزیه و تحلیل داده‌ها، امروزه داده‌کاوی در طیف وسیعی از دامنه‌های مختلف برنامه‌های کاربردی در دنیای واقعی استفاده می‌شود [۷۱، ۸۳،۹]. یک یا چند وظیفه داده‌کاوی، تکنیک‌ها و روش‌ها در این برنامه‌ها اعمال می‌شود [۴۸،۹۲] برنامه‌های کاربردی مختلف زندگی واقعی از داده‌کاوی در بخش‌های زیر ارائه شده است.

داده‌کاوی در حریم خصوصی

با ارتقا استراتژی دیجیتال، صنایع کلان داده به موتورهای جدیدی برای توسعه اقتصادی و اجتماعی تبدیل شده اند و این در حالی است که تهدیدات و خطرات امنیتی نیز در حال افزایش است. روش‌های داده کاوی حفظ حریم خصوصی نقش مهمی در کاوش، تجزیه و تحلیل حجم عظیم داده ها دارد. در این مقاله آخرین فن آوری های حفظ حریم خصوصی، از جمله فناوری اعوجاج داده، فناوری رمزگذاری داده و فناوری انتشار محدود، معرفی می شود [۹۷].

یانگ و همکاران در سه جنبه ارتباط داده کاوی با حفظ حریم خصوصی را بررسی کردند

۱. تکنولوژی تعریف حریم خصوصی

۲. تکنولوژی محاسبه حریم خصوصی

۳. تکنولوژی سنجش حریم خصوصی

تکنولوژی حریم خصوصی کاربردهای زیادی در زمینه داده‌کاوی دارد. جدول ۸ به معرفی برخی تحقیقات در خصوص حفاظت از حریم خصوصی می‌پردازد.

جدول ۸ کاربردهای حریم خصوصی در استخراج داده‌ها [۹۷]

نام روش	شرح راه حل	مزیت	عیب
DiffID3	ترکیب شاخص مکانیسم برای تولید یک درخت تصمیم ID3 برای تجزیه و تحلیل داده ها	پیاده سازی ساده و طبقه بندی با دقت بالا	کنترل تخصیص بودجه حریم خصوصی مشکل است.
DP-Bayes	اضافه کردن نویز به پارامترهای دسته‌بندی ساده بیزی برای حریم خصوصی	ساختار ساده مدل	با تکیه بر فرضیات مربوط به توزیع داده ها، هرچه مجموعه داده ها کوچکتر باشد، نویز بالاتر خواهد بود
Output perturbation mechanism	پس از آموزش مدل، اضافه کردن نویز به مدل برای حفظ حریم خصوصی	کاربرد بالا و دقت طبقه بندی در مجموعه داده های کوچکتر	ضرایب SVM باعث ایجاد خطاهای قابل توجهی می شود پردازش مجموعه داده های گسترده تر باعث چالش می شود
SuLQ-K-Means	ارسال مقادیر تخمینی برای مرکز خوشه ای و تعداد رکوردها	قابلیت محافظت از موقعیت مرکز خوشه	مقدار نویز محاسبه شده توسط حساسیت زیاد است، که در دسترس بودن نتایج خوشه بندی را کاهش می دهد
Distributed training method	تزریق نویز به گرادیان پارامترها برای محافظت از حریم	توانایی مقاومت در برابر حملات سمی	مصرف غیرضروری بودجه حریم خصوصی

ISSN: 2745-3673
http://www.Rcsj.ir

	خصوصی در شبکه های عصبی		
Moments Accountant	پیگیری هزینه های حریم خصوصی و اعمال محافظت از حریم خصوصی مناسب	قابلیت محافظت کافی از اطلاعات حساس موجود در مجموعه آموزش	بودجه خصوصی متناسب با تعداد دوره‌های آموزشی و تعداد پارامترهای مشترک جمع‌آوری می‌شود.

بخش مخابرات

داده کاوی توسط ارائه دهندگان خدمات مخابراتی و سیار مورد استفاده قرار می‌گیرد تا استراتژی‌هایی را تدوین و طراحی کنند مانند (۱) کمپین بازاریابی (۲) نگهداری مشتری (۳) بسته هایی برای مشتریان مستقر در بخش‌بندی مشتریان(۴) استفاده بهینه از زیرساخت های ارتباطی و غیره. با استفاده از طبقه‌بندی و خوشه‌بندی، ارائه دهندگان خدمات موبایل می‌توانند استراتژی‌های خود را برای کمپین بازاریابی خود تنظیم کنند. با کمک خوشه‌بندی و به دنبال آن طبقه بندی، مشتریان را می توان به انواع مختلف گروه‌ها تقسیم کرد و حرکات آنها را پیش‌بینی نمود.

استراتژی‌ها و بسته‌های خاص بازاریابی را می‌توان تدوین کرد تا بتوان با پیش‌بینی حرکت‌های مشتریان آنها را حفظ کرد. براساس گروه‌های مشتری تعریف شده بسته‌های خاصی نیز می‌توانند بر اساس نیازهای گروه‌های مختلف مشتریان فرموله شوند. برای طراحی بسته ها، می‌توان تجزیه و تحلیل انجمنی استفاده کرد. الگوی استفاده از شبکه را می توان برای شناسایی استفاده کم و یا بیش از حد زیرساخت های شبکه به کمک داده کاوی تجزیه و تحلیل کرد تا زیرساخت‌ها بتوانند به صورت بهینه مورد استفادقرارگیرند [۴۸، ۹۲،۲۵،۶۲].

بخش خرده فروشی

صاحبان بخش خرده فروشی و سوپرمارکت‌ها می‌توانند از مزایای داده‌کاوی بهره‌مند شوند. با کمک داده‌کاوی می‌توانند (۱) پیش بینی رفتار خرید مشتریان (۲) تجزیه و تحلیل سبد خرید (۳) انتخاب مشتریان (۴) قرار دادن محصولات در قفسه‌ها (۵) ارائه پیشنهادات موثر-کوپن تخفیف (۶) تقسیم بندی مشتری و غیره را جهت کشف رفتار خرید مشتریان و تجزیه و تحلیل سبد خرید به کمک قوانین انجمنی انجام دهند. با استفاده از انجمن‌ها، اقلام پرتکرار براساس سطح پشتیبانی و اطمینان داده را می توان از داده های فروش کشف کرد که این مجموعه های پرتکرار را می‌توان در نزدیکی قرار داد به طوری که فروش افزایش یابد. کمپین بازاریابی می‌تواند با استفاده از گروه‌بندی RFM[41] (تازگی، تکرار، پول) طراحی شود.

به کمک خوشه بندی می‌توان به تجزیه و تحلیل داده‌های فروش پرداخت و بهترین مکان به عنوان مثال قفسه را برای قرار دادن محصول در نظر گرفت و بهترین پیشنهادات را ارائه داد به طوری که فروش افزایش یابد. داده‌های فروش را نیز می‌توان برای استخراج بخش‌بندی‌های مختلف مشتریان به کمک خوشه بندی و یا طبقه بندی تحلیل نمود.کمپین‌های بازاریابی متفاوت و تبلیغات - پیشنهادات را می توان برای مشتریان بخش‌بندی شده سفارشی کرد. رفتار با مشتری که با دفعات کمتر خرید می - کند اما مقدار زیادی می‌خرد متفاوت از مشتری است که با دفعات زیاد اما مقادیر کمتر خرید می‌کند [۴۸،۹۲،۲۵،۶۲].

تجزیه و تحلیل داده های مالی

داده کاوی در صنعت مالی و بانکی، تجزیه و تحلیل داده‌های سیستماتیک و داده های مالی را تسهیل می‌کند. داده‌کاوی برای تجزیه و تحلیل داده‌های مالی می‌تواند برای (۱) پیش بینی پرداخت وام (۲) تجزیه و تحلیل سیاست اعتباری مشتری (۳) تقسیم بندی مشتریان برای بازاریابی هدفمند (۴)کشف پولشویی و سایر جرایم مالی و غیره استفاده شود.

41 RFM :RECENCY,FREQUENCY,MONETARY

ISSN: 2745-3673
http://www.Rcsj.ir

با کمک رتبه‌بندی و انتخاب ویژگی‌ها، روش داده‌کاوی، تاریخچه پرداخت مشتری می‌تواند برای کشف (۱) تاریخ اعتبار (۲) پرداخت به نسبت درآمد (۳) مدت وام مشتریان و غیره تحلیل شود. این پیش‌بینی به بانک‌ها و موسسات مالی در تصمیم - گیری در مورد سیاست اعطای وام و نیز اعطای وام به مشتریان طبق امتیاز آنها کمک خواهد کرد. حالا این روزها، بانک ها و موسسات مالی نمره سیبیل[۴۲] را بررسی می کنند، که بر اساس داده کاوی، از مشتریان قبل اعطای وام به آنها است [۴۸،۹۲،۲۵،۶۲].

بخش بهداشت و درمان

اخیراً، داده‌کاوی به طور گسترده‌ای در مراقبت‌های بهداشتی برای (۱) شناسایی و تجزیه و تحلیل بیماری های مزمن (۲) شناسایی و کشف علائم، علل احتمالی و داروها برای درمان‌های موثر (۳) ردیابی مناطق ریسک بالا و مستعد شیوع بیماری (۴) طراحی برنامه برای کاهش شیوع بیماری (۵) شناسایی مناطق بیماران و غیره استفاده می‌شود. در بخش مراقبت‌های بهداشتی، در تصویربرداری و تست‌های آزمایشگاهی داده‌های آزمایشات آزمایشگاه و گزارش‌ها با استفاده از عملیات داده‌کاوی مانند خوشه- بندی، طبقه‌بندی، انجمن‌ها و تشخیص نقاط پرت تجزیه و تحلیل می‌شوند. این عملیات برای شناسایی، کشف، پیش‌بینی علائم بیماری‌های مزمن استفاده می‌شود: علائم، علل احتمالی و داروها به منظور درمان موثر بیماری تجزیه و تحلیل می‌شود. این تحلیل‌ها را می‌توان برای شناسایی و پیگیری بیشتر مناطق پر خطر که مستعد شیوع بیماری هستند گسترش داد. بر اساس این تجزیه و تحلیل‌ها، کمپین‌ها می‌توانند برای مناطق مختلف طراحی شوند تا مردم را از بیماری آگاه سازند و هشدار دهند. با استفاده از داده‌کاوی، مقایسه مداوم علائم، علل و داروها، تجزیه و تحلیل داده ها را می‌توان تا ساختن درمان‌های موثر و شناسایی عوارض جانبی انجام داد[۴۸،۹۲،۲۵،۶۲].

کشف تقلب و پیشگیری از جرم

داده‌های پرت را می‌توان در مقدار زیادی داده با استفاده از داده‌کاوی استخراج کرد. داده‌های پرت می‌توانند به کمک استخراج الگوهای بدون تکرار در داده‌ها شناسایی شوند. الگوهای بدون تکرار به طور کلی متعلق به فعالیت‌های کلاهبرداری و جنایی است. از این رو، به کمک تشخیص داده‌های پرت و یا استخراج الگوی نادر، تقلب های احتمالی را می توان شناسایی و پیش بینی کرد به طوری که وقوع جرایم قابل پیشگیری است [۴۸،۹۲].

مدیریت ارتباط با مشتری[۴۳]

روابط خوب مشتریان با جذب مشتریان مناسب‌تر ایجاد د و منجر به و حفظ و نگهداری بهتر می‌شود. داده‌کاوی با تقویت CRM به تعریف و پیش بینی (۱) بازاریابی پایگاه داده (۲) جذب مشتری و کمپین های حفظ مشتری و غیره کمک می‌کند [۴۸،۹۲].

فمینا و همکاران نیز مدلی برای پیش‌بینی رفتار مشتریان جهت ارتقا فرآیندهای تصمیم‌گیری برای حفظ مشتریان با ارزش پیشنهاد نمودند. در این مدل یک چارچوب داده‌کاوی کارآمد برای مدیریت ارتباط با مشتری ارائه شده‌است و دو مدل دسته‌بندی ساده بیزی و شبکه عصبی مورد مطالعه قرار گرفته‌است تا نشان داده شود که دقت شبکه عصبی نسبتاً بهتر است [۸].

[42] CIBIL: Credit Information Bureau Limited
[43] CRM: Customer Relationship Management

ISSN: 2745-3673
http://www.Rcsj.ir

سیستم های توصیه‌گر [۴۴]

سیستم های توصیه‌گر پیشنهادات و توصیه‌های متنوعی را به ذینفعان می‌دهد که ممکن است استفاده از داده‌کاوی را برای کاربران جالب کند. سیستم‌های پیشنهاددهنده معاملات کاربران، پروفایل‌های کاربران، کلمات کلیدی، ویژگی‌های مشترک در میان اقلام برای تخمین یک مورد برای کاربر را بررسی می‌کنند.

بسیاری از تکنیک‌های استخراج داده‌ها مانند یادگیری ماشین، آمار، بازیابی اطلاعات و غیره در سیستم های توصیه کننده استفاده می‌شود. به عنوان مثال، در بازاریابی، سیستم پیشنهادی ممکن است موارد زیادی را پیشنهاد کند که یا مشابه موارد درخواست کاربر در گذشته است یا جزء ترجیحات دیگر کاربرانی است که سلیقه مشابه کاربر دارد[۲۲].

بازاریابی آنلاین / تجارت الکترونیکی

فروشندگان برندهای مختلف بازاریابی آنلاین و تجارت الکترونیکی نیز از داده کاوی برای ارتقا داده های کسب و کارخود استفاده می‌کنند. برای مثال: (۱) فروشندگان تجارت الکترونیکی کمترین قیمت محصول را با استفاده از متن‌کاوی در وب کشف می‌کنند، (۲) فروشندگان فست فودهای بزرگ زنجیره‌ای الگوی سفارش مشتریان، زمان انتظار، اندازه سفارشات و غیره با کاوش داده‌های بزرگ برای ارتقا تجارب مشتریان خود بدست می‌آورند (۳) ارائه دهندگان خدمات رسانه‌های آنلاین نیز از داده‌کاوی برای فهمیدن اینکه چگونگی یک سریال یا فیلم در میان مشتریان محبوب می‌شود، استفاده می‌کنند[۲۲].

داده‌کاوی در صنایع هوایی

حوزه هوانوردی همیشه در جستجوی راه های جدید برای بهبود ایمنی است. با این حال، با توجه به مقادیر زیادی از داده - های حمل و نقل هوایی که روزانه جمع آوری می‌شوند، تجزیه و تحلیل این حجم داده به صورت دستی غیر ممکن می‌شود.با زمینه نسبتاً جدیدی از داده کاوی، ما قادر به تجزیه و تحلیل از طریق شناسایی مقادیر غیر مجاز داده‌ها برای یافتن الگوها و ناهنجاری‌ها هستیم که رویدادهای بالقوه را قبل از وقوع آن‌ها نشان می‌دهند. روش‌های داده کاوی مشخص‌شده در این حوزه عبارتند از: الگوریتم‌های یادگیری چندگانه، مدل‌های مارکوف پنهان، مدل‌های نیمه مارکوف پنهان، و پردازش زبان طبیعی [۸۹].

داده‌کاوی در حوزه سلامت

مرجع شماره [۹۴] ساختار معماری و چالش های تجزیه و تحلیل داده های مراقبت های بهداشتی را در معرض دید قرار داده است. در مطالعه دیگری [۸۵] اهمیت امنیت و مسائل در اجرای موفقیت آمیز سیستم های مراقبت های بهداشتی نشان داده شده است. بل و همکاران [۱۰] در مورد نقش کلان داده ها در بهبود کیفیت خدمات مراقبت با جمع آوری و پردازش حجم زیادی از داده های تولید شده توسط سیستم های مراقبت سلامت بحث می کنند. در [۱۰۸] تکنیک های داده کاوی برای تجزیه و تحلیل مراقبت های بهداشتی ارائه شده است، به ویژه آنهایی که در برنامه های مراقبت های بهداشتی مورد استفاده قرار می گیرند مانند تحلیل بقا و شباهتهای بیماری.مطالعه [۱۲] یک چارچوب تجزیه و تحلیل کلان داده های مراقبت بهداشتی برای حمایت از منابع چند بعدی کلان داده های مراقبت های بهداشتی پیشنهاد داده است. هدف از این چارچوب، تجزیه و تحلیل حجم داده ها با استفاده از روش های داده کاوی است.

- انواع تجزیه و تحلیل داده در حوزه سلامت

به طور کلی تجزیه و تحلیل داده در حوزه سلامت به چهار دسته تقسیم می‌شود :

تجزیه و تحلیل تشخیصی، پیش بینی ,تجویزی و توصیفی؛

44.Recommender System

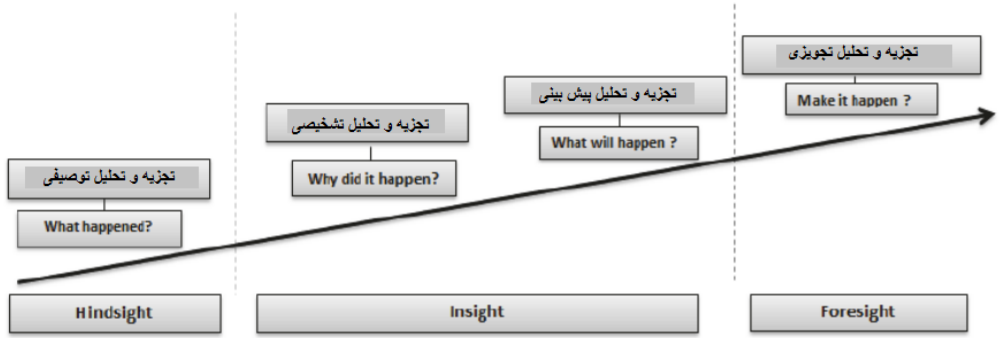
ISSN: 2745-3673
http://www.Rcsj.ir

تجزیه و تحلیل توصیفی : شامل توصیف وضعیت فعلی و گزارش دادن آنها است. برای انجام این تجزیه و تحلیل چندین تکنیک استفاده می شود. به عنوان مثال ابزار آمار توصیفی مانند هیستوگرام ها و نمودارها از تکنیک های تجزیه و تحلیل توصیفی هستند.

تجزیه و تحلیل تشخیصی : توضیح می دهد که چرا رخدادهای خاص رخ داده است و عوامل ایجاد کننده آنها چیست. به عنوان مثال، تجزیه و تحلیل تشخیصی تلاش می کند تا دلایل بهبودی برخی از بیماران را با استفاده از چندین روش مانند خوشه بندی و درخت تصمیم گیری مشخص کند.

تجزیه و تحلیل پیش بینی : نشان دهنده توانایی پیش بینی وقایع آینده است. همچنین در شناسایی روندها و تعیین احتمالهای نتایج نامطلوب کمک می کند. به عنوان مثال پیش بینی کند که آیا بیمار تحت تاثیر عوارض جانبی دارو قرار بگیرد یا نه. مدل‌های پیش‌بینی شده اغلب با استفاده از تکنیک های یادگیری ماشین ساخته می شوند.

تجزیه و تحلیل تجویزی : هدف آن پیشنهاد اقدامات مناسب جهت تصمیم گیری مطلوب است. به عنوان مثال، تجزیه و تحلیل تجویزی ممکن است پیشنهاد یک درمان داده شده را که احتمال عوارض جانبی بالایی داشته باشد را رد کند. درختان تصمیم‌گیری و شبیه‌سازی مونت کارلو نمونه‌هایی از روش‌های انجام شده برای انجام تحلیلی تجویزی هستند.شکل ۱ مرحله‌های تجزیه و تحلیل برای حوزه بهداشت و درمان را نشان می‌دهد.



شکل ۱ تجزیه و تحلیل حوزه سلامت [۲۸]

جمع‌بندی و نتیجه گیری

برای تجزیه و تحلیل کارآمد حجم زیادی از داده‌ها جهت کشف دانش از آن، نیاز به تکامل داده‌کاوی است. دامنه‌های برنامه‌های کاربردی داده کاوی نیز به طور منظم افزایش می‌یابد از این رو، یافتن الگوریتم و روش‌های یکپارچه‌ای که می‌توانند روی برنامه‌های کاربردی مختلف با یا بدون کمترین تغییرات پیاده‌سازی شوند، مورد نیاز است.

جدول ۶ چالش های مربوط به پژوهش های داده کاوی

چالش‌ها	آدرس دهی شده بوسیله
توسعه نظریه یکپارچه سازی داده کاوی	Jackson (2002) Padhy et al. (2012)
استفاده از رابط ها و عوامل هوشمند	Padhy et al. (2012)
توسعه سیستم تطبیقی،مقاوم در برابر خطا و گسترش پذیر	Sawant et al. (2013)
توانایی تغییر مداوم و ارائه درک جدید	Liao et al. (2012)
ادغام روشهای کیفی و کمی	Liao (2007)
یادگیری متریک فاصله برای داده‌های بزرگ	Wang and Sun (2014)
فرمول بندی معیار عمومی یادگیری از راه دور	Wang and Sun (2014)
بهبود کارایی و مقیاس پذیری الگوریتم های داده کاوی	Han et al. (2012)
پرداختن به انواع داده های متنوع	Han et al. (2012)

ISSN: 2745-3673
http://www.Rcsj.ir

	Silva et al. (2013)
تعامل کاربر	Han et al. (2012) Esling and Agon (2012); Geng and Hamilton (2006)
داده کاوی داده های مکعب گرا چند بعدی	Han et al. (2012)
فراآموزی برای انتخاب یا ترکیب خودکار اقدامات مناسب	Geng and Hamilton (2006)
رویکردهای همگرایی و ترکیبی	Esling and Agon (2012)
داده کاوی بدون پارامتر	Esling and Agon (2012)
معیار جامع	Esling and Agon (2012)
پویایی الگوریتم استخراج تطبیقی	Esling and Agon(2012)
جدا کردن، حفظ حریم خصوصی و استخراج تدریجی	Ceglar and Roddick (2006)
کاوش تعاملی و تکراری	Ceglar and Roddick (2006)
خوشه بندی داده های XML	Algergawy et al. (2011)
تعادل بین مقیاس پذیری و کیفیت الگوریتم های خوشه بندی	Algergawy et al. (2011)
تشخیص تکامل توزیع داده ها	Silva et al. (2013)
انسجام ساختار خوشه	Carpineto et al. (2009)
تکنیک های بصری سازی پیشرفته برای ارائه نمای کلی بهتر با نتایج خوشه ای	Carpineto et al. (2009)
کاهش زمان یادگیری / آموزش	Gibert et al. (2010)
اقدامات جایگزین برای کیفیت خوشه در یادگیری بدون نظارت و نیمه نظارت شده	Campello et al. (2015)
کاوش الگوی نادر در جریان داده ها	Koh and Ravana (2016)
قابلیت استخراج الگوی نادر به صورت بلادرنگ	Koh and Ravana (2016)
استخراج الگوی نادر در مجموعه داده های احتمالی	Koh and Ravana (2016)
استخراج الگوهای پویا و نماینده در جریان داده ها برای الگوهای مکرر غیرقطعی	Lee and Yun (2017)

اکثر سیستم‌های داده‌کاوی ترکیبی از انواع روش‌های مدیریت انواع مختلف داده‌ای، وظایف داده‌کاوی و حوزه‌های کاربردی را به کار می‌گیرند [۳۵]. وظایف و تکنیک‌های مختلف مورداستفاده در داده‌کاوی به شرکت ها برای موارد مختلفی کمک می‌کند مانند (۱) کسب دانش، و (۲) افزایش سودآوری با اعمال اصلاحاتی در عملیات و رویه‌ها، (۳) کاهش ریسک تصمیم‌گیری از طریق تجزیه و تحلیل الگوها و روندهای پنهان[۲۲].

تعدادی از چالش‌های تحقیقات داده‌کاوی توسط بسیاری از محققان بیان شده است [۱۲۱]. برخی از آنها در جدول ۹ ارائه شده و نیاز به توجه بیشتری دارد. سرانجام، نتیجه‌گیری می‌شود که (۱) یکسان سازی و یکپارچگی، مقیاس‌پذیری و بهینه‌سازی الگوریتم‌ها و روش‌های داده کاوی، (۲)داده کاوی چند بعدی مکعب گرا و (۳) کاوش بلادرنگ مقیاس پذیر حوزه‌هایی از داده کاوی است که نیاز به توجه بیشتر محققان دارد.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] Abuaiadah D (2015) Using bisect k-means clustering technique in the analysis of arabic documents. ACM Trans Asian LowResour Lang Inf Process 15(3):1–17
[2] Algergawy A, Mesiti M, Nayak R, Saake G (2011) XML data clustering: an overview. ACM Comput Surv 43(4):1–25
[3] Anderberg MR (2014) Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks, vol 19. Academic Press, USA
[4] Angiulli F, Fassetti F (2013) Exploiting domain knowledge to detect outliers. Data Min Knowl Discov 28(2):519–568
[5] Angiulli F, Fassetti F (2016) Toward generalizing the unification with statistical outliers: the gradient outlier factor measure. ACM Trans Knowl Discov Data 10(3):1–26
[6] Akulwar, P., Pardeshi,S., and Kamble.A.,(2018), "Survey on Different Data Mining Techniques for Prediction," 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2nd International Conference on, Palladam, India, pp. 513-519.
[7] Arora RK, Gupta MK (2017) e-Governance using data warehousing and data mining. Int J Comput Appl 169(8):28–31
[8] Bahari, F.T. & Elayidom, M.S. (2015). An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour. Procedia Computer Science. 46. 725-731. 10.1016/j.procs.2015.02.136.
[9] Baker RSJ (2010) Data mining for education. In: McGaw B, Peterson P, Baker E (eds) International encyclopedia of education, 3rd edn. Elsevier, Oxford, UK.
[10] Belle.A., Thiagarajan,R.,. Soroushmehr,SMR, Navidi,F., Beard, D. A. and Najarian,K.,(2015), "Big data analytics in healthcare," BioMed Research International, vol. 2015, Article ID 370194.
[11] Bhatnagar V, Ahuja S, Kaur S (2015) Discriminant analysisbased cluster ensemble. Int J Data Min Modell Manag 7(2):83–107
[12] Bochicchio,M. ,Cuzzocrea,A. and Vaira.L.(2016), "A big data analytics framework for supporting multidimensional mining over big healthcare data," in Proceedings of the 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016, pp. 508–513, usa
[13] Bouguessa M (2013) Clustering categorical data in projected spaces. Data Min Knowl Discov 29(1):3–38
[14] Campello RJGB, Moulavi D, Zimek A, Sander J (2015) Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Trans Knowl Discov Data 10(1):1–51
[15] Carpineto C, Osinski S, Romano G, Weiss D (2009) A survey of web clustering engines. ACM Comput. Surv. 41(3):1–38
[16] Ceglar A, Roddick JF (2006) Association mining. ACM Comput Surv 38(2):1–42 28. Chen YL, Weng CH (2009) Mining fuzzy association rules from questionnaire data. Knowl Based Syst 22(1):46–56
[17] Chen M, Han J, Yu PS (1996) Data mining: an overview from a database perspective. IEEE Trans Knowl Data Eng 8(6):866–883
[18] Chan P. K, Salvatore JS (1997) On the accuracy of metalearning for scalable data mining. J Intell Inf Syst 8:5–28 96.
[19] Chandra P, Gupta MK (2018) Comprehensive survey on data warehousing research. Int J Inform Technol 10(2):217–224
[20] Chen YL, Weng CH (2009) Mining fuzzy association rules from questionnaire data. Knowl Based Syst 22(1):46–56
[21] Craw S., Wiratunga N., Rowe R. C. (2006) Learning adaptation knowledge to improve case-based reasoning. Artif Intell 170:1175–1192
[22] Data mining examples: most common applications of data mining (2019). https://www.softwaretestinghelp.com/datamining-examples/. Accessed 27 Dec 2019

ISSN: 2745-3673
http://www.Rcsj.ir

[23] Data mining—applications & trends. https://www.tutorialspoint.com/data_mining/dm_applications_trends.htm
[24] Das R, Kalita J, Bhattacharya (2011) A pattern matching approach for clustering gene expression data. Int J Data Min Model Manag 3(2):130–149
[25] Devi SVSG (2013) Applications and trends in data mining. Orient J Comput Sci Technol 6(4):413–419
[26] Dincer E (2006) The k-means algorithm in data mining and an application in medicine. Kocaeli Univesity, Kocaeli
[27] Du.G, Ruhe G.. (2014) Two machine-learning techniques for mining solutions of the ReleasePlannerTM decision support system. Inf Sci 259:474–489
[28] El Aboudi, N. & Benhlima, L. (2018). Big Data Management for Healthcare Systems: Architecture, Requirements, and Implementation. Advances in Bioinformatics. 2018. 1-10. 10.1155/2018/4059018.
[29] Enke D., Thawornwong S.(2005) The use of data mining and neural networks for forecasting stock market returns. Expert Syst Appl 29:927–940
[30] Esling P, Agon C (2012) Time-series data mining. ACM Comput Surv 45(1):1–34
[31] Fan C.Y., Fan P.Sh., Chan T.Y., Chang Sh.H. (2012) Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. Expert Syst Appl 39:8844–8851
[32] Fayadd U, Piatesky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AAAI Press/The MIT Press, Massachusetts Institute of Technology. ISBN 0–262 56097–6 Fayap
[33] Fayadd U, Piatesky-Shapiro G, Smyth P (1996) Knowledge discovery and data mining: towards a unifying framework. In: Proceedings of the 2nd ACM international conference on knowledge discovery and data mining (KDD), Portland, pp 82–88
[34] Friedman JH (1997) Data mining and statistics: What is the connection? in: Keynote Speech of the 29th Symposium on the Interface: Computing Science and Statistics, Houston, TX, 1997
[35] Fu Y (1997) Data mining: tasks, techniques, and applications. IEEE Potentials 16(4):18–20
[36] Geng L, Hamilton HJ (2006) Interestingness measures for data mining: a survey. ACM Comput Surv 38(3):1–32
[37] Gheware SD, Kejkar AS, Tondare SM (2014) Data mining: tasks, tools, techniques and applications. Int J Adv Res Comput Commun Eng 3(10):8095–8098
[38] Gibert K, Sanchez-Marre M, Codina V (2010) Choosing the right data mining technique: classification of methods and intelligent recommendation. In: International Congress on Environment Modelling and Software Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
[39] Gupta A, Mumick IS (1995) Maintenance of materialized views: problems, techniques, and applications. IEEE Data Eng Bull 18(2):3
[40] Gupta MK, Chandra P (2019) MP-K-means: modified partition based cluster initialization method for k-means algorithm. Int J Recent Technol Eng 8(4):1140–1148
[41] Gupta MK, Chandra P (2019) HYBCIM: hypercube based cluster initialization method for k-means. IJ Innov Technol Explor Eng 8(10):3584–3587. https://doi.org/10.35940/ijitee. j9774.0881019
[42] Gupta MK, Chandra P (2019) A comparative study of clustering algorithms. In: Proceedings of the 13th INDIACom-2019; IEEE Conference ID: 461816; 6th International Conference on ''Computing for Sustainable Global Development''
[43] Gupta MK, Chandra P (2019) An efficient approach for selection of initial cluster centroids for k-means clustering algorithm. In: Proceedings international conference on recent developments in science engineering and technology (REDSET-2019), November 15–16

ISSN: 2745-3673
http://www.Rcsj.ir

[44] Gupta MK, Chandra P (2019) P-k-means: k-means using partition based cluster initialization method. In: Proceedings of the international conference on advancements in computing and management (ICACM 2019), Elsevier SSRN, pp 567–573
[45] Gupta MK, Chandra P (2019) An empirical evaluation of k-means clustering algorithm using different distance/similarity metrics. In: Proceedings of the international conference on emerging trends in information technology (ICETIT-2019), emerging trends in information technology, LNEE 605 pp 884–892
[46] Gupta, M.K. & Chandra, P. (2020). A comprehensive survey of data mining. International Journal of Information Technology. 1-15. 10.1007/s41870-020-00427-7.
[47] Han J, Fu Y (1996) Exploration of the power of attribute-oriented induction in data mining. Adv Knowl Discov Data Min. AAAI/MIT Press, pp 399-421
[48] Han J, Kamber M, Pei J (2012) Data mining concepts and techniques, 3rd edn. Elsevier, Netherlands
[49] Hea Z, Xua X, Huangb JZ, Denga S (2004) Mining class outliers: concepts, algorithms and applications in CRM. Expert Syst Appl 27(4):681e97
[50] Heckerman D (1998) A tutorial on learning with Bayesian networks. Learning in graphical models. Springer, Netherlands, pp 301–354
[51] Heikki M (1996) Data mining: machine learning, statistics, and databases. In: SSDBM '96: proceedings of the eighth international conference on scientific and statistical database management, June 1996, pp 2–9
[52] Hu¨llermeier E. (2005) Fuzzy methods in machine learning and data mining: status and prospects. Fuzzy Sets Syst 156:387–406
[53] Hullermeier E. (2011) Fuzzy sets in machine learning and data mining. Appl Soft Comput 11:1493–1505
[54] Hung LN, Thu TNT, Nguyen GC (2015) An efficient algorithm in mining frequent itemsets with weights over data stream using tree data structure. IJ Intell Syst Appl 12:23–31
[55] Huang Mu-Jung, Tsou Yee-Lin, Lee Show-Chin (2006) Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge. Knowl Based Syst 19:396–403
[56] Hung LN, Thu TNT (2016) Mining frequent itemsets with weights over data stream using inverted matrix. IJ Inf Technol Comput Sci 10:63–71
[57] Jackson J (2002) Data mining: a conceptual overview. Commun Assoc Inf Syst 8:267–296
[58] Jain N, Srivastava V (2013) Data mining techniques: a survey paper. Int J Res Eng Technol 2(11):116–119
[59] Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput. Surv 31(3):1–60
[60] Jin H, Wang S, Zhou Q, Li Y (2014) An improved method for density-based clustering. Int J Data Min Model Manag 6(4):347–368
[61] Kanevski M, Parkin R, Pozdnukhov A, Timonin V, Maignan M, Demyanov V, Canu S (2004) Environmental data mining and modelling based on machine learning algorithms and geostatistics. Environ Model Softw 19:845–855.
[62] Keles¸ MK (2017) An overview: the impact of data mining applications on various sectors. Tech J 11(3):128–132
[63] Khandare A, Alvi AS (2017) Performance analysis of improved clustering algorithm on real and synthetic data. IJ Comput Netw Inf Secur 10:57–65
[64] Kiranmai B, Damodaram A (2014) A review on evaluation measures for data mining tasks. Int J Eng Comput Sci 3(7):7217–7220
[65] Koh YS, Ravana SD (2016) Unsupervised rare pattern mining: a survey. ACM Trans Knowl Discov Data 10(4):1–29
[66] Kosina P, Gama J (2015) Very fast decision rules for classification in data streams. Data Min Knowl Discov 29(1):168–202

ISSN: 2745-3673
http://www.Rcsj.ir

[67] Kotsiantis SB (2007) Supervised machine learning: a review of classification techniques. Informatica 31:249–268
[68] Kumar D, Bezdek JC, Rajasegarar S, Palaniswami M, Leckie C, Chan J, Gubbi J (2016) Adaptive cluster tendency visualization and anomaly detection for streaming data. ACM Trans Knowl Discov Data 11(2):1–24
[69] Kuzey C, Uyar A, Delen (2014) The impact of multinationality on firm value: a comparative analysis of machine learning techniques. Decis Support Syst 59:127–142
[70] Lee G, Yun U (2017) A new efficient approach for mining uncertain frequent patterns using minimum data structure without false positives. Future Gener Comput Syst 68:89–110
[71] Lew A, Mauch H (2006) Introduction to data mining and its applications. Springer, Berlin
[72] Liao TW, Triantaphyllou E (2007) Recent advances in data mining of enterprise data: algorithms and applications. World Scientific Publishing, Singapore, pp 111–145
[73] Liao SH, Chu PH, Hsiao PY (2012) Data mining techniques and applications—a decade review from 2000 to 2011. Expert Syst Appl 39:11303–11311
[74] Li G, Zaki MJ (2015) Sampling frequent and minimal boolean patterns: theory and application in classification. Data Min Knowl Discov 30(1):181–225. https://doi.org/10.1007/s10618- 015-0409-y
[75] Mabroukeh NR, Ezeife CI (2010) A taxonomy of sequential pattern mining algorithms. ACM Comput Surv 43:1
[76] Maheshwari, A. K.(2015), Business intelligence and data mining. New York, NY : Business Expert Press.
[77] Mampaey M, Vreeken J (2011) Summarizing categorical data by clustering attributes. Data Min Knowl Discov 26(1):130–173
[78] Menardi G, Torelli N (2012) Training and assessing classification rules with imbalanced data. Data Min Knowl Discov 28(1):4–28. https://doi.org/10.1007/s10618-012-0295-5
[79] Mezyk E., Unold O. (2011) Machine learning approach to model sport training. Comput Hum Behav 27:1499–1506
[80] Mihoci A (2017) Modelling limit order book volume covariance structures. In: Hokimoto T (ed) Advances in statistical methodologies and their application to real problems. IntechOpen, Croatia. https://doi.org/10.5772/66152
[81] Moawia,E. Yahia ME, El-taher ME (2010) A new approach for evaluation of data mining techniques. Int J Comput Sci Issues 7(5):181–186
[82] Morik K, Bhaduri K, Kargupta H (2011) Introduction to data mining for sustainability. Data Min Knowl Discov 24(2):311–324
[83] Mukherjee S, Shaw R, Haldar N, Changdar S (2015) A survey of data mining applications and techniques. Int J Comput Sci Inf Technol 6(5):4663–4666
[84] Mukhopadhyay A, Maulik U, Bandyopadhyay S (2015) A survey of multiobjective evolutionary clustering. ACM Comput Surv 47(4):1–46
[85] Olaronke.I., Oluwaseun,O., (2016), Bigatan ealthcare:
[86] Prospects, challenges and resolutions," in Proceedings of the 2016 Future Technologies Conference, FTC 2016, pp. 1152–1157, usa.
[87] Padhy N, Mishra P, Panigrahi R (2012) A survey of data mining applications and future scope. Int J Comput Sci Eng Inf Technol 2(3):43–58
[88] Padhraic S (2000) Data mining: analysis on grand scale. Stat Method Med Res 9(4):309–327. https://doi.org/10.1191/ 096228000701555181
[89] Pagels, D. A.(2015), "Aviation Data Mining," Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal, vol. 2(1), Article 3.
[90] Pei Y, Fern XZ, Tjahja TV, Rosales R (2016) 'Comparing clustering with pairwise and relative constraints: a unified framework. ACM Trans Knowl Discov Data 11:2
[91] Politano PM, Walton RO (2017) Statistics & research methodol. Lulu. com
[92] Ponniah P (2001) Data warehousing fundamentals. Wiley, USA
[93] Rafalak M, Deja M, Wierzbicki A, Nielek R, Kakol M (2016) Web content classification using distributions of subjective quality evaluations. ACM Trans Web 10:4

ISSN: 2745-3673
http://www.Rcsj.ir

[94] Raghupathi,W., Raghupathi,V.,(2014), "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, vol. 2, article 3
[95] Reddy D, Jana PK (2014) A new clustering algorithm based on Voronoi diagram. Int J Data Min Model Manag 6(1):49–64
[96] Rustogi S, Sharma M, Morwal S (2017) Improved Parallel Apriori Algorithm for Multi-cores. IJ Inf Technol Comput Sci 4:18–23
[97] Saranya,K., Premalatha,K., and Rajasekar, S.S.,(2015), "A survey on privacy preserving data mining," 2015 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, pp. 1740-1744, doi: 10.1109/ECS.2015.7124885.
[98] Saeed S, Ali M (2012) Privacy-preserving back-propagation and extreme learning machine algorithms. Data Knowl Eng 79–80:40–61
[99] Sawant V, Shah K (2013) A review of distributed data mining using agents. Int J Adv Technol Eng Res 3(5):27–33.
[100] Shah-Hosseini H (2013) Improving K-means clustering algorithm with the intelligent water drops (IWD) algorithm. Int J Data Min Model Manag 5(4):301–317
[101] Sharma M (2014) Data mining: a literature survey. Int J Emerg Res Manag Technol 3(2):1–4
[102] Silva JA, Faria ER, Barros RC, Hruschka ER, de Carvalho ACPLF, Gama J (2013) Data stream clustering: a survey. ACM Comput Surv 46(1):1–31
[103] Silva A, Antunes C (2014) Multi-relational pattern mining over data streams. Data Min Knowl Discov 29(6):1783–1814. https:// doi.org/10.1007/s10618-014-0394-6
[104] Singh Y, Bhatia PK, Sangwan OP (2007) A review of studies on machine learning techniques. Int J Comput Sci Secur 1(1):70–84
[105] Sim K, Gopalkrishnan V, Zimek A, Cong G (2012) A survey on enhanced subspace clustering. Data Min Knowl Discov 26(2):332–397
[106] Smith Kate A, Gupta Jatinder ND (2000) Neural networks in business: techniques and applications for the operations researcher. Comput Oper Res 27:1023–1044
[107] Sohrabi MK, Roshani R (2017) Frequent itemset mining using cellular learning automata. Comput Hum Behav 68:244–253
[108] Sun,J., Reddy,C.K.(2013), "Big data analytics for healthcare," in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 1525, Chicago, Ill, USA
[109] Tan KC, Teoh EJ, Yua Q, Goh KC (2009) A hybrid evolutionary algorithm for attribute selection in data mining. Expert Syst Appl 36(4):8616–8630
[110] Tew C, Giraud-Carrier C, Tanner K, Burton S (2013) Behaviorbased clustering and analysis of interestingness measures for association rule mining. Data Min Knowl Discov 28(4):1004–1045
[111] Thompson B (2004) Exploratory and confirmatory factor analysis: understanding concepts and applications. American Psychological Association, Washington, DC (ISBN:1-59147-093-5)
[112] Top 14 useful applications for data mining. https://bigdatamadesimple.com/14-useful-applications-of-data-mining/. Accessed 20 Aug 2014
[113] Tsai Ch.-F., Hsu Y.-F., Lin Ch.-Y., Lin W.-Y.g (2009) Intrusion detection by machine learning: a review. Expert Syst Appl 36:11994–12000
[114] Turban E, Aronson JE, Liang TP, Sharda R (2007) Decision support and business intelligence systems. 8th edn, Pearson Education, UK
[115] Venkatadri M, Reddy LC (2011) A review on data mining from past to the future. Int J Comput Appl 15(7):19–22
[116] Wang B, Rahal I, Dong A (2011) Parallel hierarchical clustering using weighted confidence affinity. Int J Data Min Model Manag 3(2):110–129
[117] Wang F, Sun J (2014) Survey on distance metric learning and dimensionality reduction in data mining. Data Min Knowl Discov 29:534–564
[118] Wang L, Dong M (2015) Exemplar-based low-rank matrix decomposition for data clustering. Data Min Knowl Discov 29:324–357

ISSN: 2745-3673
http://www.Rcsj.ir

[119] Weiss SH, Indurkhya N (1998) Predictive data mining: a practical guide. Morgan Kaufmann Publishers, San Francisco
[120] Wetherill GB (1987) Regression analysis with application. Chapman & Hall Ltd, UK
[121] Yang Q, Wu X (2006) 10 challenging problems in data mining research. Int J Inf Technol Decis Making 5(4):597–604
[122] Zacharis NZ (2018) Classification and regression trees (CART) for predictive modeling in blended learning. IJ Intell Syst Appl 3:1–9
[123] Zhang W, Li R, Feng D, Chernikov A, Chrisochoides N, Osgood C, Ji S (2015) Evolutionary soft co-clustering: formulations, algorithms, and applications. Data Min Knowl Discov 29:765–791