اکتشاف دانش و داده‌کاوی در پژوهش‌های کمّی و کیفی: مقایسه روش‌شناسی‌های شبکه‌های عصبی مصنوعی (ANN) و نظریه بنیانی(GT)

محمود قاضی طباطبایی*، خدیجه طیّاری، ابوعلی ودادهیر*****

(تاریخ دریافت: ۱۳۸۸/۹/۱۲ ، تاریخ تصویب: ۱۳۸۸/۱۲/۱۸)

چکیده

در سال‌های اخیر، شاهد حرکتی مستمر از واکاوی‌ها، پژوهش‌ها و پردازش‌های صرفاً نظری و روش‌محور به پژوهش‌ها و پردازش‌های داده‌محوریم، حرکتی که به نحو احسن خود را در ظهور و توسعه روش‌های اکتشاف دانش، به ویژه داده‌کاوی و فنون خاص آن، نشان داده است. به‌رغم تصوّر رایج، داده‌کاوی صرفاً به پژوهش‌های کمّی و آماری محدود نمی‌شود و در پژوهش‌های کیفی هم شاهد ظهور تحوّلات مشابهی بوده‌ایم. در این مقاله با فرض تطبیق‌پذیری اکتشاف دانش و داده‌کاوی در پژوهش‌های کمّی و کیفی، به طور مشخص روش داده‌کاویِ شبکه‌های عصبی مصنوعی را به مثابه رویکردی نوین در پردازش چندمتغیرهٔ داده‌ها و اطلاعات و به مثابه رویکردی در حال ظهور و گسترش در روش‌های آنالیز چندمتغیرهٔ آماری، و روش داده‌کاویِ نظریهٔ بنیانی را در مدیریت و تحلیل داده‌های کیفیِ مقایسه کرده و وجوه تمایز و اشتراک آنها را بیان می‌کنیم. در این مقاله نشان داده‌ایم که صرف نظر از وجوه متمایز دو روش‌شناسیِ داده‌کاوی از حیث پارادایم، خاستگاه و فرایندهای اکتشاف و پردازش و نوع داده، هر دو روش‌شناسی از ماهیّت و رویکردی پسینی، چند رشته‌ای و میان‌رشته‌ای، استقرایی، اکتشافی، فرایندمحور، داده‌محور[1]، انعطاف‌پذیر و معطوف به رابطه (رابطه‌مدار) بین هستارها و مقوله‌ها بهره می‌برند.

واژگان اصلی: داده‌کاوی، اکتشاف دانش، روش‌شناسی، شبکه‌های عصبی مصنوعی (ANN)، نظریهٔ بنیانی(GT)

* دانشیار گروه جمعیت‌شناسی دانشکده علوم اجتماعی دانشگاه تهران، smghazi@ut.ac.ir
** کارشناس ارشد آمار زیستی و پژوهشگر و متخصصِ مدیریت پایگاه‌های داده و اطلاعات، taiyarie@gmail.com
*** استادیار گروه انسان‌شناسی دانشکده علوم اجتماعی دانشگاه تهران، vedadha@ut.ac.ir

[1] . Data-driven

مقدمه

در اغلب اجتماعات علمی و دانشگاهی جهان، تا یکی‌دو دهه قبل آموزشِ روش‌ها و فنون پژوهش و پردازش داده‌ها و اطلاعات در رشته‌های متفاوت علوم، از جمله علوم انسانی و اجتماعی، بر رویکردها و روش‌های روش‌محور[۲] و/ یا نظریه‌محور[۳] استوار بود. در دانشگاه‌ها و دانشکده‌های علوم کاربردی بخش عمده‌ای از فرایند آموزش به معرفی روش‌ها، فنون یا کاراکترها و کاربردهای خاص آنها و/ یا به معرفی و توسعهٔ نظریه‌ها، مدل‌ها و روش‌های نظریه‌محوری اختصاص داشت که از معیارهای "نیکویی برازش"[۴] بالاتری برخوردار بودند. در این رویکرد، پژوهشگران و استادان دانشگاه توجه اندکی به ماهیّت داده، انواع و ویژگی‌های آن و رویکردها و استراتژی‌های داده-محور[۵] مبذول می‌داشتند.

ظهور و گسترش کامپیوترهای مین‌فرم[۶] در اواسط دهه ۱۹۷۰ و تکامل آن به کامپیوترهای شخصی (PC) در سال‌ها و دهه‌های بعد، به مثابه نخستین انقلاب یا تحوّل پارادایمی در حوزه روش‌ها و فنون کمّی پژوهش و پردازش داده‌ها و اطلاعات، نه فقط در عمل چیزی از غفلت و کم‌توجهی پژوهشگران و دانشمندان به داده و ماهیّت و ویژگی‌های آن کم نکرد، بلکه حتی تمرکز آنان را به توسعه و درست‌آزمایی بیشتر روش‌های متعارف پژوهش و تحلیل، به ویژه از نوع تحلیل‌های چندمتغیرهٔ آماری، سوق داد. در حقیقت، این تحوّل پارادایمی با حذف و/ یا تعدیل بسیاری از موانع و محدودیت‌های شناختی، کاربردی و عملیاتی، توان محاسباتی و کامپیوتری آماردانان، پردازشگران و پژوهشگران را به طرز نظرگیری ارتقا بخشید. آموزش‌های دانشگاهی مسلّط در این دوره همچنان به شکل متمرکز بر رویکردها و روش‌های روش‌محور و/ یا نظریه‌محور استوار بود.

"عصر اطلاعات"[۷] را می‌توان نیروی محرکه یا موتور شکل‌گیری "انقلاب" یا "تحوّل پارادایمی دوّم" در حوزه روش‌ها و فنون پژوهش و پردازش داده‌ها و اطلاعات قلمداد کرد. انقلاب و انفجار اطلاعاتی، عمدتاً با خاستگاهی خارج از آکادمیا، در حقیقت با ایجاد فضا یا زمینه‌ای جدید و متفاوت برای تعریف و کاربرد روش‌ها و فنون پژوهش و پردازش داده‌ها و اطلاعات به نحو دیگری پژوهشگران این حوزه را به چالش کشید (به نقل از هِیر و همکاران، ۲۰۰۹). در حالی که پیش از این تحوّل پژوهشگران عادت داشتند داده‌هایی را پردازش و تفسیر کنند که مشخصاً به منظور پاسخ به سوال یا فرضیهٔ پژوهشی خاصی گردآوری شده بود، در شرایط جدید برای پژوهشگران امکان دسترسی به، و پردازش بانک‌های داده و اطلاعات پهن‌دامنه‌ای وجود دارد که صدها هزار و حتی میلیون‌ها مشاهده (به شکل مورد یا کاراکتر) را شامل می‌شوند. به عبارت بهتر، چنین به نظر می‌رسد که در این رویکرد، هم برای اجتماعات علمی و دانشگاهی و هم برای اجتماعات و سازمان‌هایی خارج از آکادمیا روزگار کاربست نمونه‌های محدود و استنتاج و تعمیم نتایجی از نمونه‌هایی با حجم کوچک به سر آمده است و پژوهشگران به بانک‌های داده و اطلاعات کافی و پهن-دامنه‌ای با دامنه موضوعی متنوع و گسترده‌ای دسترسی دارند. به همین سبب، امروزه بسیاری از سازمان‌های دانشگاهی و غیردانشگاهی به دنبال روش‌ها و استراتژی‌های نوینی‌اند که از طریق آنها به ذخیره، کشف و کاوش داده‌ها و اطلاعات اقدام کنند و یا بانک‌های داده و اطلاعات موجودشان را به دانشی کاربردی، سودمند و ارزشمند بدل کنند. بنابراین، تحوّل پارادایمی دوّم در خصوص روش‌ها و استراتژی‌های پژوهش کمّی و پردازش داده‌ها و اطلاعات یا در حقیقت نهضتی برای "رجوع به داده"[۸] و/ یا "گوش فرادادن به آنچه داده می‌گوید"[۹] را می‌توان با سه کاراکتر یا ویژگی بارز زیر نشان داد:

۱) اهمیت‌دادن و توجه جدّی سازمان‌های درون و بیرون از آکادمی به تولید دانش از طریق طراحی، تشکیل و "ذخیرهٔ داده‌ها و اطلاعات"[۱۰].

---

2. Method-driven
3. Theory-driven
4. Goodness of Fit
5. Data-driven
6. Mainframe
7. Information age
8. Return to the data
9. Let the data talk
10. Data & information warehousing

۲) توسعه و کاربرد روزافزون داده‌کاوی و روش‌ها و فنون خاص آن و سایر رویکردها و روش‌های داده‌محور و جایگزین.
۳) به‌چالش‌کشیدن رویکردها و روش‌های سنتی، متعارف و نسبتاً استاتیک داده‌پردازی و مدیریت اطلاعات، به ویژه با به پرسش کشیدن روش‌ها و تکنیک‌های دو یا چندمتغیرهٔ کلاسیک آماری یا روش‌های مبتنی بر استنباط آماری (هِیر و همکاران، ۲۰۰۹).

با توجه به تحوّلات پارادایمی فوق در حوزه داده‌پردازی و مدیریت اطلاعات، این مقاله بر داده‌کاوی در پژوهش‌ها و پردازش‌های کمّی و کیفی در علوم انسانی و اجتماعی متمرکز است. به رغم تصوّر رایج، از نظر ما اکتشاف دانش و داده‌کاوی صرفاً به قلمرو پژوهش و پردازش در داده‌های کمّی محدود نمی‌شود و در پژوهش‌های کیفی هم شاهد ظهور تحوّلات و رویکردهای مشابهی هستیم. این مقاله با اتخاذ رویکردی فراگیر و با فرض تطبیق‌پذیری روش‌های اکتشاف داده و داده‌کاوی در پژوهش‌ها و پردازش‌های کمّی و کیفی، مشخصاًیکی از روش‌های رایج در داده‌کاوی کمّی، یعنی شبکه‌های عصبی مصنوعی (ANNs)، را به مثابه رویکردی نوین و در حال ظهور در پردازش داده‌ها و اطلاعات کمّی و در آنالیز چندمتغیره آماری، با روش‌شناسی نظریهٔ بنیانی (GT) به مثابه رایج‌ترین روش داده‌کاوی در مدیریت و تحلیل داده‌های کیفی مقایسه و وجوه تمایز و اشتراک آنها را بیان می‌کند. نگاهی به بدنهٔ دانش و پیشینهٔ داده‌کاوی‌های کیفی و کمّی نشان می‌دهد این دو حوزه کم‌وبیش مستقل از هم رشد پیدا کرده و در محافل و منابع علمی معرفی شده‌اند، در حالی که این مقاله به شکل نوآورانه‌ای این دو را در کنار هم قرار داده و مقایسه و تحلیل می‌کند. مقاله حاضر نشان می‌دهد که صرف‌نظر از وجوه متمایز دو روش‌شناسی داده‌کاوی از حیث پارادایم، خاستگاه و فرایندها و استراتژی‌های آمایش و اکتشاف، پردازش و پایش (رصد) داده‌ها و اطلاعات، هر دو روش‌شناسی از ماهیّت و رویکردی پسینی، استقرایی، اکتشافی، داده-محور، فرایندمحور، انعطاف‌پذیر و معطوف به رابطه (رابطه‌مدار) بین هستارها و مقوله‌ها بهره می‌برند.

این مقاله در چهار بخش تنظیم شده است. ابتدا داده‌کاوی تعریف شده و دربارهٔ ماهیّت، ویژگی‌ها و فنون اصلی آن خلاصه‌وار بحث شده است. سپس، روش‌شناسی شبکه‌های عصبی مصنوعی به مثابه پارادایم یا رویکردی نو در داده‌کاوی کمّی و مبتنی بر "هوش مصنوعی"[11] معرفی می‌شود. در ادامه و به طور خلاصه، دربارهٔ روش‌شناسی داده‌کاوی کیفی نظریه بنیانی در مدیریت و پردازش داده‌های کیفی بحث شده است و در نهایت، این دو رویکرد روش‌شناختی در کاوش داده‌های کمّی و کیفی بررسی و مقایسه شده‌اند.

داده‌کاوی: رویکردی نو برای برگردانِ داده به دانش

در برداشتی متوسط، داده‌کاوی فرایند تحلیل داده از چشم‌اندازها یا زوایای گوناگون و تلخیص و تبدیل آن به دانش و/ یا اطلاعاتی سودمند تعریف می‌شود. به عبارت دیگر، داده‌کاوی یعنی فرایند کشف ویا استنتاج الگوهای بالقوه سودمند، اطلاعات معتبر و بدیع، دانش پنهان و قابل‌فهم موجود در داده‌ها ویا بانک (پایگاه) داده‌ها[12] (فیاد و همکاران، a ۱۹۹۶؛ بانج و جادسون، ۲۰۰۵؛ هادسن و کوهن، ۲۰۰۰). فرایند داده‌کاوی اطلاعات و دانشی را در اختیار شما قرار می‌دهد که افراد برای تصمیم‌گیری هوشمندانه دربارهٔ مشکلات و مسائل پیچیده و چندلایهٔ پیش روی خود به آنها نیاز دارند. داده‌کاوی راه‌حلی اساسی برای عصر" انفجار داده‌ها[13]" است؛ عصری که از سویی در داده‌ها غرق هستیم و از سوی دیگر تشنه دانشی سودمند و کاربردی هستیم. منطق و ضرورت توسعه و کاربرد مضاعف روش‌های داده‌کاوی به محدودیت‌ها و ویژگی‌های رویکردها و روش‌های سنتی برگردان داده به دانش برمی‌گردد. مهم‌ترین ویژگی روش‌های سنتی یا کلاسیک این است که بر تحلیل‌ها و تفاسیر یدی[14] استوارند (فیاد و همکاران، a۱۹۹۶). بدین معنی که رویکردهای کلاسیک نسبت به تحلیل داده، اساساً بر یک یا چند تحلیلگر مبتنی است که با داده آشنایند و در مقام واسط

---

11. Artificial intelligence
12. Databases
13. The age of data explosion
14. Manual Analysis & Interpretation

بین داده و کاربران و فرآورده‌ها عمل می‌کنند. بنابراین، شکل یدیِ بررسی یک مجموعهٔ داده بسیار بطئی، پرهزینه و قائم به طرز تفکر شخصی است. در حالی که تودهٔ داده‌ها در حال افزایش باشد، شکل یدی مدیریت داده‌ها در بسیاری از حوزه‌ها عملی نیست. پایگاه یا بانک داده‌ها به لحاظ حجمی[۱۵] به دو طریق در حال بسط و فزونی است:

- افزایش در شمار موارد ثبت شده ویا ابژه در بانک داده (N).
- افزایش در شمار حوزه‌ها یا خصیصه‌های مربوط به یک مورد یا ابژه (d).

در چنین شرایطی، همهٔ رویکردها و روش‌هایی که بتوانند توانایی‌های تحلیل را برای مدیریت خرده اطلاعات کامپیوتری حجیم تقویت کنند، از جمله روش‌های داده‌کاوی، حائز اهمیت فراوان‌اند. این رویکردها و روش‌ها عمدتاً کامپیوترمحور[۱۶] بوده و با توسل به کامپیوتر به فرد کمک می‌کنند تا به نحوی مسئلهٔ امروزین تورم و اضافگی داده[۱۷] را مدیریت کند.

داده‌کاوی صرفاً یکی از اصطلاحات یا عناوینی است که برای تعریف فرایند استخراج الگوها و اطلاعات سودمند از داده به‌کار برده شده است. علاوه بر داده‌کاوی، اصطلاحات دیگری از جمله "استخراج دانش"[۱۸]، "اکتشاف اطلاعات"[۱۹]، "تحصیل اطلاعات"[۲۰] و "دیرینه‌شناسی یا تبارشناسی داده‌ها"[۲۱] برای تشریح فرایند فوق‌الذکر استفاده شده است. به علاوه، گریگوری پیاتتسکی- شاپیرو[۲۲] برای اوّلین بار در ۱۹۹۱ اصطلاح "کشف داده از پایگاه یا بانک داده‌ها (KDD)"[۲۳] را به منظور تصریح بر این امر که دانش محصول غایی گونه‌ای فرایند اکتشاف داده‌محور است، وضع کرد. از این نظر، داده‌کاوی مرحله‌ای از فرایند KDD قلمداد می‌شود (پیاتتسکی-شاپیرو، ۱۹۹۱).

از نظر بسیاری از پژوهشگران و صاحب‌نظران، از جمله اسامه فیاد و همکاران (a, b ۱۹۹۶)، در حالی که KDDفرایند کلی کشف دانش سودمند از داده را توصیف می‌کند، داده‌کاوی به مرحلهٔ خاصی از این فرایندِ فراگیر مربوط می‌شود. این مؤلفه دارای دو هدفِ اساسی است: نخست) کشف دانش یا اطلاعاتی سودمند که منعکس‌کنندهٔ ویژگی‌های عمومی داده باشد، دوّم) کشف و استخراج الگوهایی درباره داده‌های فعلی به منظور پیش‌بینی داده‌ها و شواهد آتی. داده‌کاوی در واقع کاربرد الگوریتم‌های خاصی برای استخراج الگوها از داده است. در این تعریف، داده‌کاوی مرحله مهمی از فرایند KDD است که کاربرد الگوریتم‌های خاصِ تحلیل و کشف داده را دربرمی‌گیرد که مجموعهٔ خاصی از الگوها و مدل‌های منبعث از داده را تولید می‌کند. درک واقعی این تعریف نیازمند توجه به برخی از مفاهیم و مؤلفه‌های آن است. واژهٔ فرایند در این تعریف بر این واقعیت دلالت دارد که KDD یک‌جا و به صورت مقطعی اتفاق نمی‌افتد، بلکه فرایندی چندمرحله‌ای[۲۴] است. همچنین، در این تعریف داده‌ها مجموعه نامتجانس و متنوعی از واقعیت‌ها و شواهد را دربرمی‌گیرند. از جمله انواع داده‌ها ویا بانک داده‌ها می‌توان به فایل‌های فِلت[۲۵]، بانک‌داده‌های رابطه‌ای[۲۶] - مانند جداول، موارد ثبت شده، خصیصه‌ها، مکانیزم‌ها و نمایه‌ها/ شاخص‌ها-، بانک‌داده‌های تراکنشی[۲۷]، بانک‌داده‌های متنی[۲۸] - مانند مانند گزارش‌ها، یادداشت‌ها، ایمیل‌ها، صفحات اینترنتی، داستان‌ها، خبرها، مصاحبه‌ها- بانک‌داده‌های فضایی[۲۹]، سری‌های زمانی[۳۰]،

---

15. Size
16. Computer-intensive
17. Data overload
18. Knowledge Extraction
19. Information Discovery
20. Information Harvesting
21. Data archeology
22. Piatetsky-Shapiro, Gregory
23. Knowledge Discovery in Databases (KDD)
24. Multi-stages
25. Flat files
26. Relational databases
27. Transactional databases
28. Text databases
29. Space database

بانک‌داده‌های چندرسانه‌ای[۳۱] و داده‌های متوالی[۳۲] - نظیر توالی‌های زیستی، پروتئین‌ها، DNA و جز آنها اشاره کرد. افزون بر این، در تعریف فوق، الگو[۳۳] در حکم زیرمجموعه‌ای از داده و/ یا مدلی کاربردی برای آن زیرمجموعه است. استخراج یک الگو هم تأکیدی است بر برازش یک مدل به داده، و هم فراتر از آن، به منزلهٔ نیل به هر توصیف یا تعریف سطح بالایی از یک مجموعه از داده است. الگوهای استخراج شده بایستی بدیع، دست‌کم برای سیستم و ترجیحاً برای کاربر، سودمند و قابل فهم باشند.

فرایند اکتشاف دانش و جایگاه داده‌کاوی در آن

همان‌گونه که شکل ۱ نشان می‌دهد، KDD دارای مراحل و فازهای متفاوتی است که داده‌کاوی فقط یکی از فازها یا جریان‌های اصلی آن است. فرایند KDD ماهیّتی تعاملی و از سرگیرانه (تکراری)[۳۴] دارد که مشتمل بر چندین مرحله و تصمیم‌های بسیاری است که بایستی کاربر اتخاذ کند.

از نظر برچمن و آناند (۱۹۹۶) و همچنین فیاد و همکاران (۱۹۹۶ a, b)، فرایند KDD مشتمل بر مراحل و/ یا جریان‌های بنیادین زیر است:

نخست) بسط و توسعه فهم یا ایده‌ای دربارهٔ حوزهٔ کاربرد KDDو تعیین اهداف فرایند KDD از منظر کاربر یا مصرف‌کننده.

دوّم) ایجاد یا انتخاب یک مجموعه دادهٔ هدف (Target) ویا تمرکز بر زیرمجموعه‌ای از متغیرها و یا نمونه‌هایی از داده که قرار است اکتشاف معطوف به آنها باشد.

سوّم) آمایش و پردازش مقدماتی و پالایش داده‌ها. در این مرحله، عمدتاً فعالیت‌هایی نظیر حذف ناهمواری‌ها و مسائل و اختلالات احتمالی داده؛ جمع‌آوری اطلاعات ضروری برای مدل‌بندی و تبیین این مسائل؛ تصمیم‌گیری در خصوص استراتژی‌هایی برای مدیریت داده‌های گمشده و تبیین اطلاعات مبتنی بر توالی- زمانی و فعالیت‌هایی از این قبیل صورت می‌گیرد.

چهارم) تقلیل و به‌تصویرکشیدن داده[۳۵]. تلاش برای پیداکردن ویژگی‌های سودمندی برای نمایش و تصویرسازی داده، البته با در نظر داشتن اهداف کار. در این فاز، پژوهشگر با تقلیل داده و روش‌های خاص تبدیل، شمار نظرگیری از متغیرها را تقلیل داده و به نمایش‌های نسبتاً ثابتی برای داده‌ها نائل می‌شود.

پنجم) انطباق اهداف فرایند KDD با روش داده‌کاوی خاصی، مثلاً روش‌هایی معطوف به تلخیص، طبقه‌بندی، رگرسیون، خوشه‌بندی و مواردی جز آنها.

ششم) تحلیل اکتشافی و انتخاب فرضیه و مدل. این مرحله مشتمل بر فعالیت‌هایی معطوف به انتخاب الگوریتم‌های داده‌کاوی و گزینش روش‌هایی مناسب به منظور جستجو و شناسایی الگوهای داده است.

هفتم) داده‌کاوی. در این مرحله تلاش بر این است تا در الگوهای مورد نظر بر اساس یک یا چند مجموعه مؤلفهٔ مرتبط با الگوریتم‌های خاص داده‌کاوی - از جمله روش‌های بیان یا اظهار مدل، معیارهای ارزیابی مدل و روش‌های کنکاش پارامتر و مدل و با استفاده از روش‌ها و فنون مبتنی بر هوش مصنوعی - از جمله یادگیری ماشین، شناسایی الگوها و آمار - کنکاش شود. در مجموع، در این مرحله دو نوع از اهداف دنبال می‌شود:

یک) تأیید[۳۶] (تلاش برای تأییدپذیری فرضیه‌های مورد نظر کاربر).

دو) کشف[۳۷] (تلاش سیستم برای پیدا کردن الگوهای جدید).

---

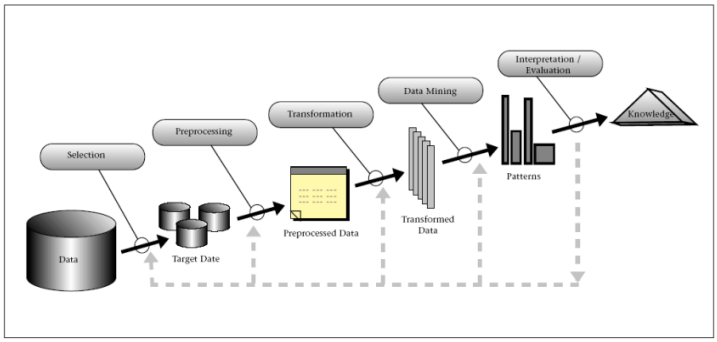
30. Time series
31. Multi-media databases
32. Sequence data
33. Pattern
34. Iterative
35. Data reduction & projection
36. Confirmation
37. Exploration

گفتنی است که هدف کشف خود به دو هدف کوچک‌تر تقسیم‌پذیر است: پیش‌بینی (سیستم‌ها بـه الگوهـایی بـرای پـیش‌بینـی رفتارهای آتی هستارهای مورد نظر نائل می‌شوند) و توصیف (تلاش سیستم‌ها برای پیدا کردن الگوهایی برای ارائـه بـه یـک کـاربر، البته به شکلی فهم‌پذیر برای انسان).

هشتم) تفسیر و ارزشیابی الگوهای استخراج‌شده. در این مرحله، برای نیل به تکرار بیشتر و/ یا به هر سبب موجّه دیگری امکـان برگشت به هر یک از مراحل هفت‌گانهٔ قبلی وجود دارد. در این مرحله فعالیت‌هایی نظیر تصویرسازی از الگوها و مدل‌های اسـتخراج شده نیز صورت می‌گیرد.

نهم) اعتبارسنجی و استفاده از دانش کشف‌شده. در این مرحله، فعالیت‌هایی در زمینهٔ دانش کسب‌شده به شکل بی‌واسطه و/ یـا از طریق سیستم‌های دانشی بزرگ‌تر و مرتبط‌تر صورت می‌گیرد. این مرحله فعالیت‌هایی نظیر آزمون و حل تضـادهای احتمـالی در بدنه دانش و/ یا مغایت آن با دانش‌های مورد باور (استخراج‌شده) قبلی را نیز دربرمی‌گیرد.

شکل ۱. پیوستاری از داده به دانش: طرحواره‌ای کلی از مراحل تشکیل‌دهندهٔ فرایند اکتشاف دانش و جایگاه داده‌کاوی در آن



منبع: برچمن و آناند، ۱۹۹۶؛ فیاد و همکاران، a ۱۹۹۶: ۴۱

نگاهی به فرایند KDD نشان می‌دهد که این فرایند می‌تواند به شکل قابل توجهی خاصیت تکرارشوندگی (از سـرگیری)[۳۸] ویـا اشتمال بر حلقه‌ها (گره‌هایی) بین مراحل داشته باشد. اهمیت خاص مرحله یا مؤلفه هفتم این فراینـد (داده‌کـاوی) بـه منزلـه کـم-اهمیتی سایر مراحل نیست، زیرا مراحل دیگر نیز دارای کاربردهای خاص خودند. وانگهی، نکته‌ای که توجه به آن ضروری است ایـن است که کاربرد کورکورانه و نامناسب روش‌های داده‌کاوی یا همان چیزی که در پیشینه علم آمار آن را "لایروبی یا زه‌کِشی داده"[۳۹] می‌نامند، می‌تواند فعالیت بسیار خطرناکی باشد، زیرا این فعالیت به سادگی می‌توانـد بـه کشـف و اسـتخراج الگوهـای بـی‌معنـی و نامعتبر منجر شود.

داده‌کاوی کمّی: شبکه‌های عصبی مصنوعی (ANNs)

چنان‌که در بخش‌های قبلی مقاله اشاره شد، فرایند KDD و مشخصاً داده‌کاوی از این قابلیت برخوردار است کـه داده‌هـا یـا بانـک داده‌های متنوع و نامتجانسی، از جمله داده‌های غیرمتنی و کمّی ارتباطی و تراکنشی، را کاوش و ارزشیابی کند. در چنین شـرایطی، داده‌کاوی ماهیّتی کمّی پیدا کرده و عمدتاً از روش‌ها و فنون کمّی بهره می‌برد. که در این شرایط از کارایی لازم برخوردارند. شبکه-های عصبی مصنوعی (ANNs) در زمرهٔ رایج‌ترین و کارآمـدترین رویکردهـا یـا روش‌هـای داده‌کـاوی اسـت. ANNs، روش‌هـای کامپیوتر-محوری[۴۰] اند که ضمن بسط دانسته‌های ما را از مغز انسان، با الهام از شبکه‌های بیولوژیکی و عملکـرد مـوازی اجـزای آن،

[38]. Iterative
[39]. Data dredging
[40]. Computer-intensive

یعنی نرون‌ها، به حل مسائل بسیار پیچیده می‌پردازند. به‌رغم آن‌که اجماعی در تعریفِ شبکه‌های عصبی مصنوعی وجود ندارد، شاید اکثر صاحب‌نظران و داده‌کاوان با تعریف ذیل موافق باشند:

"شبکه‌های عصبی مصنوعی، شبکه‌ای از واحدهای اصطلاحاً عملگر یا پردازشگرهای بسیار کوچکی‌اند که هر کدام دارای حافظهٔ محلی کوچکِ منحصر به‌فردی بوده و از طریق کانال‌هایی ارتباطی که معمولاً اعدادی را به روش‌های خاص حمل می‌کنند، با همدیگر ارتباط دارند. این واحدها فقط داده‌های محلی و ورودی‌هایی را که از طریق کانال‌ها دریافت می‌کنند، پردازش می‌کنند" (هادسن و کوهن، ۲۰۰۰).

همچنین، از نظر بانج و جادسن[41](۲۰۰۵)، "شبکهٔ عصبی، عملگری با توزیع موازی بسیارگسترده است که ویژگی ذاتی آن ذخیره‌سازی اطلاعات تجربی و آماده‌سازی آنها برای استفاده است."به باور آنها، شبکهٔ عصبی مغز انسان را از دو نظر شبیه‌سازی می‌کند:

- شبکه اطلاعات را در خلال فرایند یادگیری به‌دست می‌آورد.
- از قدرت ارتباطات بین‌نرونی،که به سیناپس‌ها معروف‌اند، برای ذخیره‌سازی اطلاعات استفاده می‌شود" (بانج و جادسن، ۲۰۰۵).

شماری از ANNها که تاکنون ساخته شده و برای حل مسائل گوناگون به‌کار رفته‌اند، مدل‌ها یا ماکت‌هایی از شبکه‌های عصبی بیولوژیکی بوده‌اند. با وجود این، شبکه‌های عصبی صرفاً از نوع بیولوژیک نیستند، برای این‌که بخش عمده‌ای از انگیزه‌ها و مساعی دانشمندان در ساختن شبکه‌های عصبی از علاقه خاص آنها به خلق و تولید سیستم‌های مصنوعی و اقدامات پژوهشی آنها برای محاسباتی شبیه آنچه دائماً در مغز انسان صورت می‌گیرد، نشأت می‌گیرد.

ویژگی‌های شبکه عصبی مصنوعی

ANNها با وجود این‌که در بسیاری از موارد با سیستم‌های عصبی طبیعی ویا بیولوژیک قابل مقایسه نیستند، دارای ویژگی‌هایی‌اند که آنها را در برخی شرایط و کاربردها از روش‌های سنتی پردازش داده‌ها – مثلاً رویکردها یا روش‌های آماری کلاسیک و/ یا مبتنی بر پیش‌فرض‌هایِ توزیعِ نمونه‌گیری) متمایز می‌کند. این ویژگی‌ها عبارت‌اند از

- قابلیت یادگیری
- پراکندگی اطلاعات
- قابلیت تعمیم
- پردازش موازی
- مقاوم بودن در برابر خطاها و تورش‌ها.

از آنجای که یک ANN از مجموعه‌ای از نرون‌ها تشکیل می‌شود و نرون به لحاظ عملکرد و/ یا نوع پردازش روی داده در زمرهٔ روش‌های غیرخطی طبقه‌بندی می‌شود، عملکرد شبکه عصبی از نوع کاملاً پیچیده و غیرخطی است. به عبارت بهتر، خاصیت غیرخطی پردازشگرها (نرون‌ها) در کل شبکه عصبی توزیع شده است. وانگهی، به سبب این‌که مدل‌های ANNs مبتنی بر داده‌اند، این مدل‌ها به راحتی قادرند هر گونه تغییر احتمالی در ساختار داده‌ها را دریافت کند، و بدون ایجاد اختلال یا اغتشاشی[42] در پردازش همه نرون‌ها، پردازشگرها را با ساختار جدید منطبق کنند. این در حالی است که در روش‌های غیرخطی سنتی آماری، در صورت عدم برقراری پیش‌فرض‌های آماری، هر نوع تغییر احتمالی در داده‌ها ضمن به پرسش کشیدن صحت نتایج، به تغییر روش آماری مورد استفاده برای پردازش داده‌ها منجر می‌شود. این ویژگی ANNها را قابلیت یادگیری می‌نامند.

---

41. Bunge J.A., Judson D.H.
42. Noise

هنگامی که ANNها به پردازش داده می‌پردازند، دانش درونِ داده به تمامی عناصر شبکه آموزش داده می‌شود و رابطهٔ یک-به‌یکی بین اجزای داده و ارتباط بین نرون‌ها برقرار نمی‌شود. به تعبیر بهتر، دانشی که در ارتباط‌های بین نرون‌ها، به مثابه پردازشگرهای کوچک، وجود دارد با کل شبکه مرتبط است و به یک نرون یا پردازشگر خاص محدود نمی‌شود. بر اساس این ویژگیِ ANN که به پراکندگی اطلاعات/ پردازش اطلاعات به صورت زمینه‌ای[۴۳] معروف است، چنانچه بخشی از سلول‌های شبکه حذف شوند و/ یا عملکرد اشتباهی داشته باشند، باز هم احتمال نیل به پاسخی صحیح وجود دارد. اگر چه این احتمال برای همهٔ ورودی‌ها کاهش می‌یابد، ولی برای هیچ‌یک از آنها از بین نمی‌رود.

در صورتی‌که یک ANN در نتیجهٔ پردازش داده‌ای دانش نهفته درون آن را استخراج کند، می‌توان از آن برای حل مسائل دیگر با داده‌های مشابه استفاده کرد که این ویژگی را قابلیت تعمیم یا تعمیم‌پذیری مدل‌های ANNها می‌نامند. چنان‌که در بخش‌های قبلی مقاله بیان شد، روش‌های غیرخطیANN ها از شبکه‌های عصبی بیولوژیکی الهام گرفته‌اند و طبق عملکرد شبکه‌های عصبی بیولوژیکی، پردازش درون آنها با مؤلفه‌های آن یا همان نرون‌ها به‌صورت مستقل از هم و موازی صورت می‌گیرد. این ویژگی ANNها را "پردازش موازی" می‌نامند. پردازش موازی باعث افزایش سرعت پردازش می‌شود. همچنین، در یک شبکهٔ عصبی، هر نرون مستقل عمل می‌کند و رفتار کلی شبکه، در واقع، برآیند رفتارهای محلی نرون‌های متعدد است. این ویژگی موجب می‌شود خطاهای محلی از چشم پاسخ نهایی دور بمانند. به‌عبارت دیگر، نرون‌ها در روند همکاری خطاهای محلی یکدیگر را تصحیح می‌کنند. این ویژگی باعث افزایش قابلیت مقاومت یا تحمل‌پذیری سیستم در برابر خطاها و تورش‌ها می‌شود (چِنگ و تیترینگتون، ۱۹۹۴؛ منهاج، ۱۳۸۱).

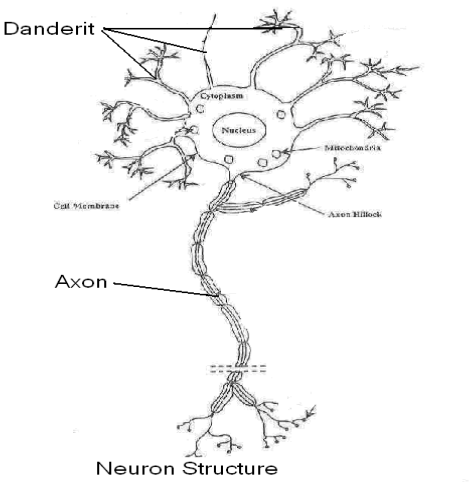
در مجموع، ویژگی‌های فوق به نحوی روش کار ANNها را نشان می‌دهند. به این ترتیب که در ANNها، ابتدا نرون‌ها به مثابه اجزای پردازشگر شبکه‌های عصبی دانش نهفتهٔ درون داده‌ها را استخراج می‌کنند و سپس با توسل به الگوی استخراج درون‌داده آنالیز داده‌ها صورت می‌گیرد. فرایند استخراج دانشِ درونِ داده در دو مرحله قابل اجر است: مرحلهٔ اول که اصطلاحاً فرایند یادگیری شبکه‌عصبی از داده نامیده می‌شود و مرحله دوّم که به آزمون شبکه عصبی موسوم است.

مؤلفه‌های و ساختارهای اصلی مدل‌های ANNها

در برداشتی کلی، شناخت شبکه‌های عصبی در گروِ شناخت مؤلفه‌ها، از جمله نرون‌ها و انواع آن، و ساختارهای شبکه‌ای آن است. چنان‌که پیش از این آمد، عنصر یا مؤلفه اصلی پردازش در مغز "نرون" نامیده می‌شود که سیگنال‌های ورودی را از طریق دندریت-های خود دریافت می‌کند و پس از پردازش آنها از طریق اکسون، خروجی را به دیگر نرون‌ها منتقل می‌کند. چنان‌که شکل ۲ نشان می-دهد، دندریت هر نرون به اکسون نرون دیگر از طریق سیناپس‌ها متصل است.

[43]. Context

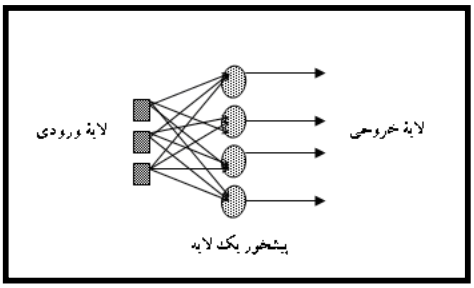
شکل ۲. مغز انسان و واحد محاسباتی آن



هر شبکهٔ بیولوژیکی در مغز از یک لایهٔ ورودی، یک لایهٔ خروجی و تعدادی لایهٔ درونی، معروف به لایهٔ پنهان یا مکنون، تشکیل شده است. هر شبکه با توجه به وظیفه‌ای که در مغز بر عهده دارد، می‌تواند شمار متفاوتی لایهٔ پنهان داشته باشد. هر کدام از این لایه‌ها تعدادی نرون دارد. در شبکه‌های بیولوژیکی، نرون‌های بسیار متفاوتی می‌توانند اطلاعات را به سیناپس گزارش دهند. نرون‌ها ورودی‌های چندگانه‌ای دریافت می‌کنند. اگر مجموع اطلاعاتی که نرون‌ها دریافت می‌کنند به یک سطح آستانه‌ای برسد، نرون‌ها آتش می‌گیرند. منتقل‌کننده‌های عصبی به شکل مؤثری قدرت یا وزن ارتباط‌ها را کنترل می‌کنند (کارتالوپولس، ۱۳۸۲). قدرت سیگنال دریافتی به وسیلهٔ نرون اساساً به کارایی سیناپس‌ها بستگی دارد. هر سیناپس در واقع یک فضای باز کوچک با ماده‌ای شیمیایی، با نام منتقل‌کنندهٔ عصبی است که آمادهٔ انتقال سیگنالی در طول فضای باز است.

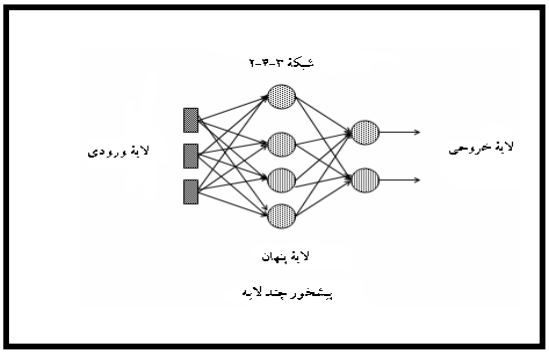
ساختارهای ANNها در سه طبقهٔ شبکه‌های تک‌لایهٔ پیش‌خور، شبکه‌های چندلایهٔ پیش‌خور و شبکه‌های بازگشتی (برگشت‌پذیر) تقسیم‌بندی می‌شوند. در این تقسیم‌بندی لایهٔ ورودی در شمارش لایه‌ها به حساب نمی‌آید. شبکه‌های تک‌لایهٔ پیش‌خور، که به آنها پرسپترون نیز می‌گویند، صرفاً یک لایهٔ خروجی دارند و لایهٔ پنهان ندارند. در این نوع شبکه‌ها، ارتباط بین نرون‌ها یک‌طرفه است و قابلیت برگشت ندارد. این شبکه‌ها معمولاً در حل مسائل خطی به‌کار گرفته می‌شوند. همچنین، این شبکه‌ها در زمرهٔ شبکه‌های ایستا طبقه‌بندی می‌شوند. شکل ۳ طرحواره‌ای کلی از ساختار شبکه‌های تک لایه را نشان می‌دهد.

شکل ۳. سیمایی کلی از شبکه‌های عصبی تک لایه



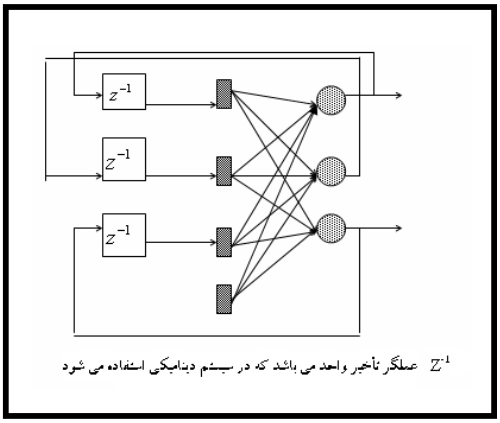
شبکه‌های پیش‌خور چندلایه همان شبکه‌های پیش‌خور تک‌لایه‌اند، با این تفاوت که در آنها یک یا چند لایهٔ پنهان وجود دارد. به این شبکه‌ها پرسپترون چندلایه نیز گفته می‌شود. شکل ۴ طرحواره‌ای کلی از ساختار شبکه‌های چندلایه با یک لایهٔ پنهان (برای نمونه، یک شبکه ۳-۴-۲) را به نمایش می‌گذارد.

شکل ۴. نمایی کلی از ساختار شبکه‌های عصبی دو لایه



شبکهٔ عصبی بازگشتی[44] آن شبکه‌هایی‌اند که در آنها، ارتباط صرفاً در یک جهت نبوده و برگشت هم دارند (شکل ۵). شایان ذکر ذکر است که در مواقعی که بین اطلاعات دینامیکی وجود داشته باشد، از این نوع شبکه‌ها استفاده می‌شود (م. ش. بیل و جکسون، ۱۳۸۰؛ شالکف، ۱۳۸۲؛ کارتالوپولس، ۱۳۸۲؛ چِنگ و تیترینگتون، ۱۹۹۴).

شکل ۵. نمایی کلی از شبکه‌های عصبی بازگشتی



نرون واحدِ پایهٔ پردازشگر اطلاعاتِ شبکهٔ عصبی است. یک نرون از سه جزء یا مؤلفه اصلی تشکیل می‌شود:

۱. مجموعه‌ای از ارتباط‌ها، با وزن‌های $w_1, w_2, \ldots w_m$، که ورودی‌های نرون را توصیف می‌کنند.

۲. تابع جمع‌کننده (ترکیب‌کنندهٔ خطی) مجموع موزون ورودی‌ها را به‌صورت $u = \sum_{j=1}^{m} w_j x_j$ محاسبه می‌کند.

۳. تابع تحریک (تابع هموارکننده) φ که خروجی نرون را به‌صورت $y = \varphi(u + b)$ محدود می‌کند.

با انتخاب نوع φ مدل نرون تعیین می‌شود. دو نوع از φ ممکن است به‌صورت زیر باشد.

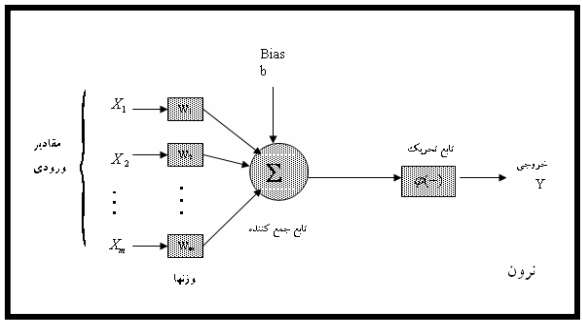
---

[44]. Recursive

- تابع گام: $\varphi(v) = \begin{cases} a & if \quad v < c \\ b. & if \quad v > c \end{cases}$
- تابع سیگموئید با پارامترهای x,y,z $\varphi(v) = z + \frac{1}{1+\exp(-xv+y)}$

شکل ۶ نمایی کلی از یک نرون را به مثابه یک واحد محاسباتی کوچک نشان می‌دهد.

شکل ۶. نمایی از نرون به مثابه واحد محاسباتی کوچک



نگاهی به پیشینه و آثار پژوهشی ANNs نشان می‌دهد که الگوریتم‌های متفاوتی برای یادگیری نرون وجود دارد. در تعریفی کلی، یادگیری یعنی برآورد پارامترهای مربوط به نرون‌های شبکه که به واسطهٔ آن، شبکه قادر به انجام کار خاصی می‌شود. دو نوع اصلی الگوریتم یادگیری، یعنی یادگیری با ناظر[۴۵] و یادگیری بدون ناظر[۴۶]، برای آموزش شبکه وجود دارد. در یادگیری با ناظر، نتایج درست (مقادیر هدف یا خروجی‌های مطلوب) معلوم بوده و در خلال آموزش در اختیار شبکه قرار می‌گیرند، به‌طوری‌که شبکهٔ عصبی بتواند وزن‌های خود را با نزدیک‌کردن خروجی‌ها به مقدارهای هدف تعدیل کند. بعد از آموزش، شبکهٔ عصبی فقط با دادن داده‌های ورودی و نظارهٔ بر چگونگی نزدیک‌شدن مقادیر خروجیِ تولیدی به مقادیر هدف آزمون می‌شود. پرسپترون، آدلاین و شبکه‌های عصبی پیش‌خور نمونه‌هایی از این شبکه‌هایند. در یادگیری بدون ناظر، اطلاعات بدون نتایج درست در اختیار شبکه قرار داده می‌شود و خود شبکه ارتباط بین داده‌ها را پیدا می‌کند. نقشه‌های خود- سازمانده[۴۷] و شبکه‌های هاپفیلد جزو این نوع شبکه‌هایند. لازم به ذکر است که در مواردی ممکن است از هر دو روش برای آموزش شبکه استفاده شود (شالکف، ۱۳۸۲).

همواره قانونی کلی در بدنه دانش و پیشینهٔ مربوط به شبکه‌های عصبی مصنوعی وجود دارد که کاربران بایستی بدان توجه کنند: "تعداد لایه‌های پنهان و تعداد نودهای روی لایه‌ها باید به حد مقبول و مناسبی باشند." شبکه‌های عصبی به دفعات بسیار فرایند آموزش را تکرار می‌کنند. بنابراین، هر چقدر تعداد لایه‌های پنهان و نودها بیشتر باشند، زمان بیشتری برای آموزش لازم است. رعایت این قانون فراگیر مزایایی دارد که از آن جمله می‌توان به اجتناب از بیش‌یادگیری (یادگیری مضاعف)[۴۸] اشاره کرد. به این قانون اصل امساک[۴۹] می‌گویند.

به طور خلاصه، در سال‌های اخیر شاهد حرکتی مستمر و فزاینده برای توسعهٔ نظری سیستم‌های دینامیکی مدل- آزاد[۵۰]، که مبتنی بر داده‌های تجربی‌اند، بوده‌ایم. ANNها جزو این دسته از سیستم‌های دینامیکی‌اند که با پردازش داده‌های تجربی و به- کارگیری روش‌های متنوع داده‌کاوی، قانون نهفته در ماورای داده‌ها را به ساختار شبکه منتقل می‌کنند. در سال‌های اخیر ANNها

---

45. Supervised learning
46. Unsupervised learning
47. Self-organizing maps (SOM)
48. Over training
49. Parsimony
50. Model free

هم از لحاظ نظری و هم به لحاظ عملی پیشرفت‌های قابل ملاحظه‌ای کرده و کاربردهای بسیاری در علوم گوناگونی مانند روانشناسی، جامعه‌شناسی، جمعیت‌شناسی، اقتصاد، بازاریابی، انواع مهندسی، زمین‌شناسی، فیزیک، اپیدمیولوژی، تغذیه، مهندسی پزشکی، پزشکی، مطالعات ریسک و مخاطره، و آمار پیدا کرده‌اند که این امر همه بسیاری از روش‌های سنتی مدل‌بندی در این علوم را به چالش کشیده است، و هم کاربران روش‌های آماری را بر آن داشته تا به کارایی و عملکرد روش ANNها را در مقام مقایسه با روش‌های کلاسیک و سنتی آماری توجه کنند. داده‌کاوی کمّی از طریق ANNها صرفاً مشتمل بر تکنیک‌های تحلیلی نیست، بلکه در واقع رویکرد یا نگاهی خاص برای آنالیز داده‌هاست. این نوع داده‌کاوی دامنهٔ متنوعی از ابزارهای تحلیلی را دربرمی‌گیرد. طرفداران این رویکرد در حال ظهور ادعا می‌کنند که در بسیاری از موقعیت‌هایی که روش‌های متعارف یا کلاسیک آماری از عهدهٔ تبیین و حل مسائل بر نمی‌آیند، می‌توان از این رویکرد یا روش‌شناسی نو کمک گرفت. به علاوه، برخلاف روش‌های متعارف آماری که محدودیت‌هایی از نوع مقیاس اندازه‌گیری یا تعداد متغیرها دارند، در داده‌کاوی کمّی به شیوه ANNها چنین محدودیت‌هایی ندارد. داده‌کاوی قادر به انعکاس مناسبات بسیار پیچیده در میان متغیرهای مورد مطالعه است. همچنین، در برداشتی کلاسیک، این نوع داده‌کاوی نیازی به پیش‌فرض‌ها یا مفروضاتی از پیش تعریف‌شده (عمدتاً پیش‌فرض‌های مبتنی بر توزیع نمونه‌گیری) ندارد، امری که در پردازش‌های آماری متعارف اجتناب‌ناپذیر به نظر می‌رسد. البته این به منزله آن نیست که داده‌کاوی با رویکرد ANNها صرفاً ماهیّتی اکتشافی[51] دارد. رویکرد ANNها در مواردی می‌تواند برای اهداف علمی از نوع تأییدی[52] نیز استفاده شود.

به اعتقاد مروّجان و کاربران شبکه‌های عصبی مصنوعی، آزاد- توزیع بودن، کامپیوتر-محوری، الهام‌گیری از شبکه‌های عصبی بیولوژیکی و عدم نیاز به پیش‌فرض‌های متعارف آماری در حل و پردازش مسائل از جمله مهم‌ترین علل تمایز و مزیت‌های این رویکرد و فنون خاص آن نسبت به سایر روش‌های پردازش چندمتغیره آماری است. به سبب همین ویژگی‌ها و شاخصه‌ها، امروزه ANNs در بسیاری از حوزه‌های علم‌فناورانه، از جمله علوم آماری و تحلیلی، علوم پزشکی، علوم اقتصادی و اجتماعی و مدیریت، کاربرد داشته و اخیراً با تلاش گروهی از دانشمندان علوم انسانی و اجتماعی برای داده‌کاوی و اقدامات معطوف به حل مسئله در این حوزه‌ها به‌کار می‌روند. بنابراین، اقدامات و پژوهش‌های علمی متعددی به منظور معرفی ماهیّت و کاربردهای ANNها و برازش مدل‌های آن در حوزه‌های متفاوت علم فناوری صورت گرفته است (م. ش. منهاج، ۱۳۸۱؛ هایکین، ۱۹۹۴؛ زورادا، ۱۹۹۲؛ افتخار و همکاران، ۲۰۰۵).

صرف نظر از همهٔ موارد فوق و ویژگی‌ها و امتیازات منسوب به ANNها نمی‌توان آن را مؤلفهٔ کاملاً مجزایی از بدنهٔ انباشته‌شدهٔ دانش و پیشینهٔ رویکردها و فنون چندمتغیرهٔ پردازش آماری دانست، برای این‌که این نوع داده‌کاوی در زمرهٔ رویکردها و روش‌های تحلیلی چندمتغیره است. همسو با سایر روش‌های پردازش چندمتغیره، ANNها نسبت به متغیرها و ورایته‌های نامناسب و نامرغوب حساس بوده و در صورت انتخاب چنین متغیرهایی، مدل‌های ANNها آسیب‌پذیر خواهند بود. (اقتباس از هِیر و همکاران، ۲۰۰۹). افزون بر این، این نوع داده‌کاوی انسان‌محور[53] بوده و اصالت با تفکر، خلاقیّت و شعور کاربر یا تحلیلگر است. بنابراین، در این رویکرد نیز هرگز به ماشین اجازه تفکر به جای انسان (کاربر یا تحلیلگر) داده نمی‌شود و این امر به ویژه در طراحی و استراتژی‌های تدوین طرح پژوهشی مطرح است. همچنین، اعتبارسنجی مواد (داده‌ها/ ورودی)، روش‌ها، فرایند و نتایج پردازش‌های صورت‌گرفته مؤلفهٔ اجتناب-ناپذیری از داده‌کاوی کمّی از طریق روش‌شناسی ANNها است. نهایتاً این‌که، همانند سایر روش‌های چندمتغیرهٔ آماری و بنابه علل متعددی از جمله ماهیت معطوف به حل مسئلهٔ مدل‌های ANN، ارزیابی اهمیت عملی و کاربردی نتایج به‌دست آمده برای این نوع داده‌کاوی حائز اهمیت فراوان است.

---

51. Exploratory
52. Confirmatory
53. Human-centered

داده‌کاوی کیفی: نظریهٔ بنیانی (GT)

چنان‌که در بخش‌های قبلی مقاله گفته شد، داده‌کاوی صرفاً با داده‌های کمّی ویا با استفاده از داده‌های عددی نیست، بلکه داده‌کاوی برای داده‌های متنی یا متن‌محور نیز قابل استفاده است. بنابراین، گزافه نخواهد بود اگر از داده‌کاوی کیفی نیز سخن به میان آوریم. برخلاف پژوهش‌ها و داده‌کاوی‌های کمّی و آماری که به سادگی می‌توان رویه‌ها و مراحل آن را از هم تفکیک کرد و تحلیل داده‌ها مرحلهٔ مستقلی از فرایند گردآوری آنهاست، در پژوهش‌های کیفی، به ویژه در روش‌شناسی "نظریه بنیانی" (نظریه برخاسته از داده)[۵۴]، مرزبندی شفافی بین مراحل و جریان‌های پژوهش وجود ندارد. به عبارت دیگر، در این رویکرد با مراحل و رویه‌های یکپارچه‌ای سروکار داریم که قویاً ماهیّت مبتنی بر تکرار (از سرگیری) دارند. به عبارت ساده‌تر، در پژوهش کیفی تحلیل داده‌ها فرایندی چندمرحله‌ای و مستمر است که دست‌کم با فرایند گردآوری داده‌ها شروع می‌شود (گلیزر و استراوس، ۱۹۶۷؛ استراوس و کوربین، ۲۰۰۸؛ کرسوِل، ۲۰۰۹). حتی برخی، نظیر جوزف مکسول،[۵۵] فراتر از این می‌روند و بر این نکته تأکید می‌کنند که تحلیل داده‌ها بخش مهمی از طرح پژوهشی اولیه است، زیرا هر مطالعهٔ کیفی نیازمند تصمیماتی درباره چگونگی تحلیل داده‌هاست. از سوی دیگر، با توجه به ماهیّت داده‌های کیفی که بیشتر به شکل متن، کلمات و عبارت‌های نوشتاری و مبتنی بر زمینه‌اند و/ یا نمادهایی که مردم، کنش‌ها و وقایع زندگی اجتماعی را توصیف می‌کنند، این داده‌ها می‌توانند حامل بیش از یک معنی باشند و برخلاف فرایند پردازش در داده‌های کمّی، پردازش و کاوش این داده‌ها کمتر حالت استاندارد داشته و تنوع قابل ملاحظه‌ای در طراحی و کاربست فرایند تحلیل داده‌های کیفی وجود دارد و این امر کم‌وبیش فرایند داده‌کاوی کیفی را پیچیده‌تر می‌کند.

نظریهٔ بنیانی مهم‌ترین رویکرد یا روش‌شناسی در داده‌کاوی کیفی است. این روش‌شناسی برای اولین بار گلیزر و استراوس[۵۶] در ۱۹۶۷ با هدف ساختن نظریه‌ای که در دل داده‌ها قرار بگیرد، مطرح کردند. این دو جامعه‌شناس با مطالعهٔ بیمارانِ در آستانهٔ مرگ و با انتشار اثر ماندگار *کشف نظریهٔ بنیانی: استراتژی‌هایی برای پژوهش کیفی*[۵۷] عملاً پایه‌های این روش‌شناسی داده‌کاوی را بنا نهادند (گلیزر و استراوس، ۱۹۶۷). این نظریهٔ بنیانی کشف نظریه از داده‌هاست که به صورت استقرایی و بر مبنای پژوهش اجتماعی تولید می‌شود. خلق نظریه بر مبنای داده‌ها و اطلاعات پژوهش به منزله آن است که مفاهیم، مقوله‌ها و فرضیه‌ها نه فقط از اطلاعات میدانی، بلکه از داده‌هایی به‌دست آمده است که در جریان پژوهش به شکل نظام‌مند تکوین یافته‌اند. از این‌رو، برخلاف رویکردهای سنتی که اساساً بر بازیافت‌پذیری نتایج اندازه‌گیری و آزمون تجربی و فرضیه و قضایای نظری تأکید می‌کنند، تأکید روش‌شناسی نظریه بنیانی بر استخراج و تولید نظریه از داده است (ذکایی، ۱۳۸۱؛ نیومن،۲۰۰۰).

داده‌کاوی کیفی از طریق نظریه بنیانی دارای سه مؤلفه اساسی است: مفاهیم[۵۸]، مقوله‌ها[۵۹] و گزاره‌ها[۶۰] - و به یک معنی فرضیه-ها بر پایه این سه مؤلفه، می‌توان سه فرایند یا استراتژی عمده برای داده‌کاوی کیفی و کاربست روش‌شناسی نظریهٔ بنیانی را در نظر گرفت:

- استراتژی‌های مفهوم‌پردازی[۶۱]، مقوله‌پردازی[۶۲] و نظریه‌پردازی[۶۳] (کدگذاری، تحلیل مضمون[۶۴]، اعتبارسنجی[۶۵] مدل نظری برخاسته از داده و جز آنها)
- استراتژی‌های زمینه‌ای شدن/ زمینه‌پردازی[۶۶] (تحلیل روایت[۶۷]، مطالعه‌های موردی فردی[۶۸] و جز آنها)

---

54. Grounded theory
55. Maxwell, Joseph A.
56. Glaser, B. G. & A. L. Strauss
57. The discovery of grounded theory: strategies for qualitative research
58. Concept
59. Category
60. Proposition
61. Conceptualization
62. Categorization
63. Theorization
64. Theme analysis
65. Validation

- استراتژی‌های نگارش و کاربرد یادداشت‌های فنی (ممو‌ها)[۶۹] و ابزارهای نمایشی (مکسول، ۱۹۹۸).

جوزف مکسول (۱۹۹۸) بر این باور است که این فرایندها یا استراتژی‌ها کاملاً از هم مستقل نبوده و قابلیت تلفیق و ترکیب دارند و پژوهشگر می‌تواند بنا بر نیاز اطلاعاتی واکاوی خود و ماهیّت سؤالات پژوهشی خود منظومه‌ای همگرا، منطبق و انعطاف‌پذیر از آنها را اتخاذ کند.

در مجموع، نظریه بنیانی نوعی رویکرد یا روش‌شناسی داده‌کاوی کیفی است که دارای ویژگی‌های اصلی زیر است:

- نظریهٔ بنیانی ماهیّتی پسینی و/ یا تبعی[۷۰] دارد. یعنی در مواقعی که نظریه‌های موجود (پیشین) توان تبیین فرایند یا شرایطی را ندارند، می‌توانیم با توسل به روش‌شناسی نظریهٔ بنیانی دربارهٔ وقوع این فرایند، شرایط یا مشکل یا افراد مورد مشاهده نظریه‌ای را صورت‌بندی کنیم.
- نظریهٔ بنیانی متناسب با موقعیت یا محیط پژوهشی است و در مقام مقایسه با نظریه‌های موجود (پیشین) در پی عرضهٔ تبیین بهتری از موقعیت نامعین است. به علاوه، این روش‌شناسی نسبت به خصوصیات فردی در زمینه یا محیط پژوهشی حساس بوده و ممکن است همهٔ پیچیدگی‌های فرایند را منعکس کند (اقتباس از: بازرگان، ۱۳۸۷).
- نظریهٔ بنیانی رویکردی مناسب برای مطالعه فرایندهای اکتسابی و/ یا برساخته‌های اجتماعی و فرهنگی[۷۱] است. مثلاً می‌توان از این رویکرد برای فهم فرایندی استفاده کرد که طی آن افراد جامعه نسبت به استعمال دخانیات و حتی مواد مخدر، به‌رغم آگاهی و شناخت نسبت به مضرات و نتایج منفی چنین تمایلاتی تمایل پیدا می‌کنند.
- روش‌شناسی نظریهٔ بنیانی از ماهیّت خوداصلاح و تکرارشونده (از سرگیرانه) برخوردار است. بدین معنی که پژوهشگر با اتکا بر پردازش مجموعه‌ای از داده‌ها سمت و سوی تحلیل‌های بعدی را هدایت و مشخص می‌کند (بازرگان، ۱۳۸۷).
- روش‌شناسی نظریهٔ بنیانی از حیث داده‌ها و منابع و همچنین فنون جمع‌آوری، مرتب‌سازی و کدگذاری داده‌ها و اطلاعات نامتجانس[۷۲] است. بدین معنی که در این روش‌شناسی محدودیتی در تکنیک‌ها و رویه‌های جمع‌آوری و مرتب-سازی و کدگذاری داده‌ها و اطلاعات وجود ندارد. به عبارت بهتر، این رویه‌ها فی‌نفسه انعطاف‌پذیر و متنوع‌اند تا مکانیکی و همگون.
- برخلاف رویکردهای سنتی در واکاوی‌های اجتماعی و رفتاری، در داده‌کاوی‌های کیفی، به ویژه روش‌شناسی نظریهٔ بنیانی، موضع پژوهشگر[۷۳] در کل جریان پژوهش منفعل، بی‌طرفانه و عاری از ارزش نیست (م. ش. بلیکی، ۲۰۰۷). به عبارت دقیق‌تر، در این روش‌شناسی شناسا و شناخته به لحاظ معرفت‌شناختی قابل تفکیک نیستند. لین ریچاردز (۲۰۰۵) بر خلاف سنت رایج در پژوهش‌های کمّی و با عنایت به مشارکت فعّال پژوهشگر در همهٔ مراحل و فرایندهای پژوهش‌های کیفی اصطلاح "ساختن/خلق داده‌ها"[۷۴] را به جای "گردآوری داده‌ها"[۷۵] برای این نوع پژوهش‌ها مناسب‌تر می‌داند (ریچاردز، ۲۰۰۵).

---

66. Contextualization
67. Narrative analysis
68. Individual case studies
69. Memoing
70. A posterior and/or post-hoc
71. Socio-cultural constructions
72. Heterogeneous
73. Researcher's stance
74. Data making
75. Data gathering

- نهایتاً این‌که، "تحلیل تطبیقی پیوسته داده‌ها"[۷۶] از جمله ویژگی‌های بارز روش‌شناسی نظریه بنیانی است. بدین ترتیب که در این روش‌شناسی، پژوهشگر ابتدا داده‌ها را گردآوری می‌کند/ می‌سازد، سپس با اتکا بر داده‌ها به فرایندهای مفهوم‌پردازی، مقوله‌پردازی و نظریه‌پردازی مشغول می‌شود. در خلال این فعالیت‌ها، پژوهشگر داده‌ها و اطلاعات اضافی لازم را گردآوری کرده و مفاهیم، مقوله‌ها و گزاره‌های حاصل از این داده‌ها و اطلاعات را با موارد قبلی، منبعث از داده‌ها و اطلاعات قبلی، مقایسه می‌کند. به این فرایند استقرایی و تدریجی تکوین مفاهیم و مقوله‌های مبتنی بر داده‌ها و اطلاعات، شیوهٔ "تحلیل مقایسه‌ای پیوسته داده‌ها" اطلاق می‌شود (بازرگان، ۱۳۸۷).

بحث و نتیجه‌گیری

این مقاله بر رویکردهای داده‌کاوی در پژوهش‌ها و پردازش‌های کمّی و کیفی در علوم انسانی و اجتماعی متمرکز است. به رغم تصوّر رایج، مقالهٔ حاضر نشان داد که اکتشاف دانش از بانک داده و داده‌کاوی صرفاً به پژوهش‌ها و داده‌های کمّی محدود نمی‌شود و در پژوهش‌های کیفی هم شاهد ظهور تحوّلات پارادایمی و رویکردهای مشابهی هستیم. این مقاله با اتخاذ رویکردی فراگیر و با فرض تطبیق‌پذیری روش‌های اکتشاف دانش و داده‌کاوی در پژوهش‌ها و پردازش‌های کمّی و کیفی، مشخصاً یکی از روش‌های رایج در داده-کاوی کمّی، یعنی شبکه‌های عصبی مصنوعی (ANNs)، را به مثابه رویکردی در حال ظهور و جدید در پردازش داده‌ها و اطلاعات کمّی و در آنالیز چندمتغیرهٔ آماری، با روش‌شناسی نظریهٔ بنیانی (GT)، به مثابه رایج‌ترین روش داده‌کاوی در مدیریت و تحلیل داده-های کیفی، مقایسه و وجوه تمایز و اشتراک آنها را بیان کرد. نگاهی به بدنهٔ دانش و پیشینهٔ داده‌کاوی‌های کیفی و کمّی نشان می‌دهد این دو حوزه کم‌وبیش مستقل از هم رشد کرده و در محافل دانشگاهی و منابع علمی معرفی شده است، در حالی که این مقاله به شکل نوآورانه‌ای این دو را در کنار هم قرار داده و مقایسه و تحلیل کرده است. مقایسهٔ دو روش‌شناسی داده‌کاوی نشان می‌دهد که صرف نظر از وجوه متمایز آنها از نظر پارادایم، خاستگاه فکری و فرایندهای اکتشاف و پردازش و تفسیر و نوع داده یا ورودی، استراتژی‌های اعتبارسنجی و نتایج هر دو روش‌شناسی بیش‌وکم از ماهیّت و رویکردی پسینی، استقرایی، میان‌رشته‌ای، اکتشافی، داده-محور، فرایندمحور، کاربردی (معطوف به حل مسئله)، انعطاف‌پذیر و معطوف به رابطه (رابطه‌مدار) بین هستارها و مقوله‌های مورد بررسی بهره می‌برند.

در هر دو روش‌شناسی هدف نیل به دانش، الگوها و نظریه‌هایی داده‌محور است و تلاش می‌کنند به آنچه از داده بیرون می‌آید یا استخراج می‌شود توجه کنند. در این مقاله، سعی شد تا اهمیت و جایگاه رویکردهای داده‌محور و داده‌کاوی در پژوهش‌ها و پردازش‌های کمّی و کیفی منعکس شود. نکته شایان توجه این است که هر دو رویکرد داده‌کاوی روش‌شناسی‌هایی برای تبدیل و برگردان داده به دانش‌اند، البته نه با شکل و رویه‌ای کلاسیک، بلکه به شکلی متمایز و در عین حال غیرهمگرا از هم. هر دو روش‌شناسی داده‌کاوی در پی نقد پردازش‌های ساده‌انگارانه، رشته‌ای، خطی، پیشینی و مبتنی بر رویکرد محض و پوزیتویستی از علم‌فناوری و روش‌های آن‌اند، البته با شکل و رویه‌ای متفاوت. این مقاله همچنین نشان داد که داده‌کاوی صرفاً منظومه یا جمعی جبری از فنون و ابزارهای پردازش نیست، بلکه رویکردی نوین برای مدیریت، برگردان، کاوش و بهره‌برداری از داده‌ها و اطلاعات است. به همین سبب، این مقاله مؤید آن است که صرف-نظر از تفاوت‌های دو رویکرد داده‌کاوی، هر دو روش‌شناسی در بسیاری از موارد، رویه‌ها و فرایندهای مشابهی را دربرمی‌گیرند و موضع نسبتاً همگرایی نسبت به داده‌ها و مدیریت داده‌ها دارند.

با توجه به تحوّلات پارادایمی در زمینهٔ ماهیّت و نوع داده‌ها، استراتژی‌های داده‌پردازی و مدیریت و اکتشاف اطلاعات، مقالهٔ حاضر این نکته اساسی را خاطر نشان می‌کند که امروزه در دانشگاه‌ها و دپارتمان‌های علوم انسانی، اجتماعی و پزشکی آموزش و معرفیِ صرفِ رویکردها، روش‌ها و فنون پژوهش و پردازش روش‌محور و/ یا نظریه‌محور و بسط و توسعه نظریه‌ها، مدل‌ها و روش‌های نظریه‌محوری که از معیارهای "نیکویی برازش" بالاتری برخوردارند، دیگر کافی به نظر نمی‌رسند، بلکه پژوهشگران و استادان

---

[76]. Constant comparative data analysis

دانشگاه باید توجه بیشتری به ماهیّت داده‌ها و خصیصه‌های آن و رویکردها و استراتژی‌های داده‌محور[۷۷] مبذول کنند. از این‌رو، این مقاله از ضرورت بازنگری جدّی و پهن‌دامنه در آموزش روش تحقیق و شیوه‌های گردآوری، آمایش (روش‌های غربالگری، کشف و هموارسازی) و پردازش و بهره‌برداری از داده‌ها و اطلاعات در علوم انسانی و اجتماعی بحث کرده و صاحب‌نظران، علاقه‌مندان، پژوهشگران و مدرّسان علوم اجتماعی و انسانی، به ویژه جامعه‌شناسان و رفتارشناسان ایرانی، را به توجه بیشتر به داده و انواع و ماهیت آن و رویکردهای داده‌کاوی و سایر روش‌های داده‌محور و ویژگی‌های مشترک و متمایز آنها دعوت می‌کند. این مقاله به سهم خود نشان می‌دهد که دانشمندان و پژوهشگران علوم اجتماعی ایران بایستی توجه و زمان و انرژی بیشتری را به شناسایی و تأمل در ماهیّت و خصوصیات داده‌ها و انبار و استخراج و تفسیر آنها اختصاص دهند. به باور ما، زمان آن فرا رسیده است که علوم اجتماعی ایران در مسیر خلق و گردآوری و انبار داده‌های حجیم و طولی و توسعه و آموزش رویکردها و فنون داده‌کاوی قدم‌های محکم‌تر و جدّی‌تری بردارد. نگاهی به تاریخ شکل‌گیری و توسعهٔ علوم اجتماعی و رفتاری نوین نشان می‌دهد که هدایت و ناوبری تحوّلات اجتماعی و مشارکت در برنامه‌های مداخله‌ای اجتماع‌محور و مهندسی اجتماعی و رفتاری و فراتر از آنها، جهش به سوی مرزهای دانش در علوم انسانی و اجتماعی بدون تولید، ذخیره، مدیریت، کاوش و استخراج و بهره‌مندی از شواهد و داده‌ها و/ یا پایگاه‌های دادهٔ حجیم و طولی و تبدیل و برگردان آن به دانشِ سودمند و در نهایت مشروعیت‌بخشی[۷۸] و تفهیم[۷۹] منطق و عقلانیت[۸۰]، سودمندی و کاربردهای آن به همهٔ گروه‌های ذینفع (اقامه‌کنندگان دعوی، اصحاب رسانه، برنامه‌ریزان، سیاستمداران، رهبران دینی، کاربران، صاحبان صنایع، دانشجویان و جز آنها) ممکن نیست.

واپسین سخن این‌که، داده‌ها و شواهدی که درباره دنیای اطراف خود کسب می‌کنیم، در واقع به مثابه شواهد و مستندات بنیادینی عمل می‌کنند که با توسل به آنها نظریه‌ها، مدل‌های نظری و تحلیلی و استراتژی‌های مقابله، شیوه‌های سازگاری و مراوده با جهانی اجتماعی را صورت‌بندی می‌شوند که در آن زندگی می‌کنیم. تردیدی نیست که نیل به این شواهد و مستندات بنیادین بدون تأسیس پایگاه‌های داده و اطلاعات معتبر و حجیم و کسب شناخت و قابلیت‌های بسنده دربارهٔ روش‌های متنوع داده‌کاوی و اکتشاف و استخراج دانش سودمند از داده‌ها امکان‌پذیر نیست.

منابع

بازرگان، عباس (۱۳۸۷) مقدمه‌ای بر روش‌های تحقیق کیفی و آمیخته: رویکردهای متداول در علوم رفتاری، تهران: نشر دیدار.

بیل، آر.، و جکسون تی (۱۳۸۰) آشنایی با شبکه‌های عصبی، ترجمه محمود البرزی، تهران: مؤسسه انتشارات علمی دانشگاه صنعتی شریف.

ذکایی، محمد سعید (۱۳۸۱) "نظریه و روش در تحقیقات کیفی"، در *فصلنامه علوم اجتماعی*، شماره ۱۷ (بهار ۱۳۸۱)، صص ۶۹-۴۱.

شالکف، رابرت جی (۱۳۸۲) شبکه‌های عصبی مصنوعی، ترجمهٔ محمود جورابیان، طناز زارع، امید استوار، اهواز: انتشارات دانشگاه چمران اهواز.

کارتالوپولس اس. وی (۱۳۸۲) منطق فازی و شبکه‌های عصبی: مفاهیم وکاربردها، ترجمهٔ محمود جورابیان، و رحمت‌اله هوشمند، اهواز: انتشارات دانشگاه چمران اهواز.

منهاج، محمد باقر (۱۳۸۱) مبانی شبکه های عصبی، تهران: انتشارات دانشگاه صنعتی امیر کبیر.

Blaikie, N. (2007) *Approaches to Social Enquiry*, 2nd Edition, Cambridge: Polity Press.

---

77. Data-driven
78. Legitimacy
79. Communicating
80. Rationale for

Brachman, R. & T. Anand (1996) "The Process of Knowledge Discovery in Databases: A Human-centered Approach" In Fayyad, U., Piatetsky-Shapiro, G., Smith, P., & R. Uthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, : 37-58.

Bunge J.A., D.H.Judson (2005), Data Mining, In Bunge, J.A., Judson, D. H.(eds.) *Encyclopedia of Social Measurement*, Elsevier Academic Press, : 617-624.

Cheng, B. & D.M. Titterington, (1994) "Neural Networks: A Review from a Statistical Perspective" *Statistical Science*, 1: 2-30.

Creswell, JW. (2009) *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*, Thousand Oaks, CA: Sage.

Eftekhar, B., Mohammad, K., Eftekhar Ardebili, H., Godsi, M., Ketabchi, E. (2005) "Comparison of Artificial Neural Network And Logistic Regression Model for Prediction of Mortality in Head Trauma Based on Initial Clinical Data" in *BMC Medical Informatics and Decision Making*, 5: 1-8.

Fay, B. (1996) *Contemporary Philosophy of Social Science*, Oxford: Blackwell Publishing.

Fayyad, U., G. Piatetsky-Shapiro & P. Smith (1996a) "From Data Mining to Knowledge Discovery in Databases", *Artificial Intelligence (AI) Magazine*, Fall: 37-54.

Fayyad, U., G. Piatetsky-Shapiro, P. Smith, & R. Uthurusamy (1996b) *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press.

Glaser, B. G., & A. L. Strauss (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*, New York: Aldine de Gruyter.

Goodman P. H. & F. H. Harrell (1998) *Neural Networks, Advantages and limitations for Biostatistical Modeling*, Available at http://www.scs.unr.edu/nevprop.

Hair, J. F., R. E. Anderson, R. l. Tatham, & W. C. Black (2009) *Multivariate Data Analysis*, 7th Edition, Prentice-hall.

Haykin, S. (1994) *Neural Networks: A Comprehensive Foundation*, New York: Macmillan.

Hudson, D. L. & M. E. Cohen (2000*) Neural Networks and Artificial* Intelligence for Biomedical Engineering, New Delhi: Prentice-Hall of India Private Limited.

Maxwell, J. A. (1998) *"Designing a Qualitative Study"* In Bickman, L. & D. J. Rog (eds.) *Handbook of Applied Social Research Methods*, Thousand Oaks, CA: Sage Publication, : 69-100.

Neuman, W. L. (2000) *Social Research Methods:Qualitative and Qualitative Approaches*, Boston: Allyn and Bacon.

Richards, M. (2005) *Handling Qualitative Data: A Practical Guide*, London: Sage Publications.

Piatetsky-Shapiro, G. (1991) "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop" in *Artificial Intelligence (AI) Magazine*, 11 (5): 68-70.

StatSoft (2009) *STATISTICA Neural Network*, Available at http://www.statsoft.com/textbook/stathome.html.

Strauss. A. & J. Corbin (2008) *Basics of Qualitative Research*, 3rd Edition, Newbury Park, CA: Sage Publications.

Zurada, J. M. (1992) *Introduction to Artificial Neural Systems*, St. Paul & New York: West Publishing Company.