مقایسهٔ آمار و داده‌کاوی

عبدالرضا ناظمی* و علی مشکانی*

چکیده

امروزه داده‌کاوی[1] مهمترین ابزار برای استفادهٔ سودمند از منابع متنوع و فراوان داده‌ها محسوب می‌شود. داده‌های دقیق در حجم بسیار زیاد و قیمت ارزان توسط شرکتها و سازمانها مختلف تولید گردیده، در بانکهای اطلاعاتی و یا داده انبارها سازماندهی می‌شوند و استفاده مؤثر از آنها توسط لایه‌های مختلف مدیریت، یک هدف عمده محسوب می‌شود. در حال حاضر، داده‌کاوی مهمترین فناوری برای بهره‌برداری مؤثراز داده‌های حجیم است و اهمیت آن رو به فزونی است. در این مقاله به بیان رابطه‌ها و تشابه‌ها و تفاوتهای آمار و داده‌کاوی، وظایف و محدودیت‌های داده‌کاوی پرداخته شده است.

واژه‌های کلیدی: آمار، داده‌کاوی، پایگاه داده‌ها[2]

مقدمه

با توجه به همه‌گیرشدن استفاده از کامپیوتر در سازمانها و شرکتهای مختلف، حجم زیادی از داده‌ها تولید می‌شود که مقدار آنها روز به روز در حال افزایش است. این موضوع سبب به‌وجود آمدن رشتهٔ جدید داده‌کاوی شده است. داده‌کاوی تلفیقی از رشته‌های آمار، علوم کامپیوتر (خصوصاً مدیریت پایگاه داده‌ها) و یادگیری ماشینی است. در زمینهٔ رابطهٔ آمار و داده‌کاوی، مقالاتی منتشر شده است. فریدمن[1] و دیوید هند[2] نیز مقالاتی در این زمینه منتشر کرده‌اند. موضوع اصلی این مقاله بررسی تفاوتهای آمار و داده‌کاوی است.

داده‌کاوی

پیشرفت فوق‌العاده در کسب و ذخیره‌سازی داده‌های عددی، سبب به وجودآمدن پایگاه داده‌های بزرگ شده است. داده‌های تبادلات تجاری، کشاورزی، ترافیک اینترنت، داده‌های نجومی، جزئیات مکالمات تلفنی، داده‌های پزشکی و درمانگاهی، مثالهایی از چنین پایگاه داده‌ها، هستند. در واقع فنون تولید و جمع‌آوری پایگاه داده‌ها بسیار سریعتر از توانایی ما در درک و استفاده از آنها رشد کرده است.

با توجه به وجود اطلاعات ارزشمند در این پایگاه داده‌ها، در دههٔ ۸۰ تلاش برای استخراج و استفاده از این اطلاعات شروع شد. به طور رسمی اصطلاح داده‌کاوی برای اولین بار توسط فیاض[3] در اولین کنفرانس بین‌المللی کشف معرفت و داده‌کاوی[4] در سال ۱۹۹۵ مطرح شد.

نگاهی به ترجمهٔ تحت اللفظی داده‌کاوی، به ما در درک بهتر این واژه کمک می‌کند. mining به معنای استخراج از منابع نهفته

* گروه آمار دانشگاه فردوسی مشهد

[1] Friedman

[2] David Hand

[3] Fayyad

[4] Knowledge Discovery and Data Mining

و با ارزش زمین اطلاق می‌شود. پیوند این کلمه با کلمهٔ داده، جستجویی عمیق برای پیداکردن اطلاعات اضافی مفید که قبلاً نهفته بودند، از داده‌های قابل دسترس حجیم، را پیشنهاد می‌کند. داده‌کاوی یک رشتهٔ نسبتاً جدید علمی است که از انجام تحقیقات در رشته‌های آمار، یادگیری ماشینی، علوم کامپیوتر، مدیریت پایگاه داده‌ها شکل گرفته‌است که مرزهای آن با مرزهای این رشته‌ها مبهم است.

داده‌کاوی دارای تعاریف متنوعی است. این تعاریف به مقدار زیادی به پیش‌زمینه‌ها ودیدگاههای افراد بستگی دارد. در اینجا ما به ارائه بعضی از این تعاریف می‌پردازیم:

۱— داده‌کاوی یک فرایند شناخت الگوهای معتبر، جدید، ذاتاً مفید و قابل فهم از داده‌ها است.(Fayyad)

۲— داده‌کاوی به فرایند استخراج اطلاعات نهفته، قابل فهم، قابل پیگیری از پایگاه داده‌های بزرگ و استفاده از آن در تصمیم‌گیریهای تجاری مهم اطلاق می‌شود.(Zekulin)

۳— داده‌کاوی، مجموعه‌ای از روشها در فرایند کشف دانش است که برای تشخیص الگوها و روابط نامعلوم در داده‌ها مورد استفاده قرار می‌گیرد.(Ferruzza)

۴— فرایند کشف الگوهای مفید از داده‌ها را داده‌کاوی گویند.(John)

تعریف کاملتری از داده‌کاوی را می‌توان به صورت زیر ارائه کرد: فرایند انتخاب، کاوش و مدل‌بندی داده‌های حجیم، برای کشف روابط نهفته با هدف به دست آوردن نتایج واضح و مفید، برای مالک پایگاه داده‌ها را، داده‌کاوی گویند.

آمار و داده‌کاوی

آمار و داده‌کاوی هر دو با روشهای تحلیل و مدل‌بندی داده‌ها مرتبط‌اند. بنابراین اشتراک زیادی بین این دو رشته وجود دارد. به عنوان یک شوخی، یکی از نویسندگان در پاسخ سئوال اینکه "داده‌کاوی چیست؟" بیان می‌کند که همان آمار است، اما با یک نام خیلی بهتر. همان طور که گفتیم یکی از رشته‌های مورد استفاده در داده‌کاوی آمار است. بر هم‌کنش این دو رشته سبب به وجود آمدن موضوعات فراوان تحقیقاتی شده و مورد علاقهٔ بسیاری از آماردانان قرار گرفته‌است.

چون تشابهات و ارتباطات بین داده‌کاوی و آمار نسبتاً معلوم هستند، ما به بررسی تفاوتهای این دو رشته می‌پردازیم.

برای یک مدت طولانی برای آماردان‌ها داده‌کاوی به عنوان مترادفی از صید داده‌ها[۵]، لایروبی داده‌ها[۶] و دستکاری داده‌ها[۷] مطرح بوده‌است. در تمامی این موارد داده‌کاوی دارای یک معنای ضمنی منفی است.

تفاوتی در نوع داده‌ها وجود دارد. آماردانان با "داده های دست اول" که برای بررسی‌کردن فرضهای خاصی جمع‌آوری و تولید شده‌اند کار می‌کنند. اما داده‌کاوها با "داده‌های دست دوم" و یا داده‌های مشاهده‌ای که اغلب از منابع مختلفی گردآوری شده‌اند، کار می‌کنند. منظور پیدا کردن وقایع مورد علاقه و اطلاعات مفیدی است که در داده‌ها مخفی شده‌اند و اغلب با اهداف ابتدایی که داده‌ها به خاطر آن جمع‌آوری شده‌اند، رابطه‌ای ندارند. از طرفی داده‌کاوی با حجم وسیعی از داده‌ها کار می‌کند و همچنین بعضی از پایگاه داده‌ها ساختار مناسب داده‌های آماری را ندارند. برای درک بهتر و کاملتر تفاوت آمار و داده‌کاوی به جدول ۱ مراجعه کنید.

وظایف و محدودیتهای داده‌کاوی

مهمترین وظایف داده‌کاوی عبارتند از

۱— آنالیز کاوشگرانه داده‌ها[۸]: با رسم جداول و نمودارها و شکلهای نموداری بدون اینکه ایده‌ای واضح از آنچه جستجو می‌کنیم، داشته‌باشیم، ما را در تحلیل داده‌ها کمک می‌کند.

۲— مدل‌بندی توصیفی: هدف شرح همه داده‌ها یا فرایند تولید داده‌هاست، فنون آن مانند برآورد توزیع، تحلیل خوشه‌ای، ...

۳— مدل‌بندی پیشگو (رده‌بندی و رگرسیون): هدف ساختن یک مدل که بتواند مقادیر متغیر پاسخ را با استفاده از متغیرهای دیگر پیشگویی کند. اگر متغیر پاسخ رسته‌ای باشد آن را رده‌بندی و اگر متغیر پاسخ پیوسته یا کمی باشد، رگرسیون گویند. تفاوت اصلی بین مدلهای پیشگو و توصیفی، در این است که، مدلهای پیشگو دارای یک متغیر یکتا هدف‌اند اما

[۵] Data fishing
[۶] Data dredging
[۷] Data snooping
[۸] Exploratory Data Analysis

جدول ۱: مقایسهٔ آمار و داده‌کاوی

آمار	داده‌کاوی
حجم داده‌ها	
کوچک و متوسط	بزرگ
نوع داده‌ها	
داده‌ها برای آزمون یک مدل یا پاسخ دادن به یک سؤال خاص جمع‌آوری شده‌اند. (داده‌های دست اول) • مطالعه‌های کنترل موردی • طرح آزمایشها • بررسی نظرخواهی • مطالعه‌های مشاهده‌ای	داده‌ها به‌طور الکترونیکی برای استفاده‌های ممکن آینده نگهداری می‌شوند. (داده‌های دست دوم) داده‌های تبادلات تجاری داده‌های ترافیک اینترنت داده‌های مکالمات تلفن داده‌های پزشکی
پردازش داده‌ها	
روشهای دستی	روشهای قویاً خودکار
پردازش داده‌ها توسط انسان به کمک کامپیوتر صورت می‌گیرد	پردازش داده‌ها توسط الگوریتم‌های کامپیوتری با کمک انسان صورت می‌گیرد
وظایف معمول	
• برازش مدل • آزمون مدل • بازه‌های اطمینان و پیش‌بینی	جستجو و شناخت الگوها رده‌بندی دسته‌بندی
اهداف تحقیق	
توسعهٔ روشهای آماری بهتر مطالعهٔ خواص آماری و ریاضی روشها	توسعهٔ الگوریتمهای بهتر و سریعتر برای اجرای وظایف مطالعهٔ عملکردهای تجربی الگوریتمهای داده‌کاوی

مدلهای توصیفی این گونه نیستند. فنون آن عبارتند از: درخت تصمیم،[۹] شبکه‌های عصبی،[۱۰] رگرسیون خطی، ...

۴— کشف الگوها و قوانین: هدف تعیین الگوها می‌باشد.

در پایان ذکر نکات زیر خالی از لطف نیست.

- داده‌کاوی یک وسیله است نه یک عصای سحرآمیز.
- داده‌کاوی نمی‌تواند، داده‌های مورد نیاز ما را تأمین کند.
- داده‌کاوی نمی‌تواند الگوهای مهم موجود در داده‌ها را به طور خودکار مشخص کند.
- حل مسائل داده‌کاوی نیاز به درک داده‌ها و معلومات در آن زمینهٔ خاص را دارد.
- تصمیم‌گیری فقط با توجه به نتیجه داده‌کاوی، عاقلانه نیست.
- روابط پیشگویی حاصل از داده‌کاوی لزوماً علت یک پدیده یا رفتار نیست.

مراجع

[1] (1995). Proceedings of the First International Conference on Knowledge Discovery and Data Mining Edited by Usama Fayyad and Ramasamy Uthurusamy Published by The AAAI Press, Menlo Park, California.

[2] Hand, D., Mannila, H. and Smyth, P. (2002). Principles of data mining, The MIT Press.

[3] Han, J. and Kamber, M. (2000). Data Mining: Concepts and Techniques, Jim Gray, series Editor, Morgan Kauafman publisher.

[4] Hand, D. J. (1999). Statistics and Data Mining: Intersecting Discipline, SIGKDD Exploration.

[5] Friedman, J. H. (1997). Data Mining and Statistics: What's the connection?, http://www-stat.stanford.edu/jhf/ftp/dm-stat.ps

[6] Giudici, P. (2003). Applied Data Mining. John Wiley and Sons.

[7] Two Crows Corporation. (1999). Introduction to Data Mining and Knowledge Discovery, Third edition.

[8] Chipman, H. (2002). Statistical learning and Data Mining, University of Waterlo.

[9] www.kdnuggets.com.

[۹] Decision Trees
[۱۰] Neural Networks