مجله گزیده‌مطالب آماری
سال ۱۸، شماره ۱، بهار و تابستان ۱۳۸۶، صص ۱۹-۱

رهنمودهایی برای بهبود کیفیت داده‌ها در طرح‌های آمارگیری با استفاده از سامانه‌های هوشمند

علی‌اصغر حائری مهریزی،[۱،*] محمدرضا یگانگی،[۲] حسین حسنی[۳]

[۱] وزارت صنایع و معادن
[۲] دانشگاه شهید چمران اهواز
[۳] دانشگاه کاردیف انگلستان

چکیده. در این مقاله، رهنمودهایی برای طراحی یک سامانه هوشمند به‌منظور بهبود کیفیت داده‌ها در تمامی مراحل تولید داده (از مرحله گردآوری تا انتشار) در یک طرح آمارگیری ارائه شده است. با استفاده از این رهنمودها ساختار سامانه‌ای هوشمند برای طرح هزینه و درامد خانوار که یکی از طرح‌های مهم مرکز آمار ایران است، پیشنهاد شده است.

۱- مقدمه

یکی از نیازهای مبرم سیاست‌مداران در تصمیم‌گیری‌های کلان هر کشور اطلاعات آماری است. بدون داشتن اطلاعات آماری کافی، جامع و به‌هنگام، نمی‌توان تصمیم صحیحی اتخاذ کرد. بر این اساس مراکز آمارگیری سالانه هزینه‌های هنگفتی را به‌منظور تهیه اطلاعات آماری مورد نیاز دولت متحمل می‌شوند.

روش‌های بسیار زیادی برای گردآوری اطلاعات آماری وجود دارند. به‌عنوان مثال می‌توان به گردآوری اطلاعات از طریق انواع پرسشنامه و به روش‌های متفاوت (با استفاده

واژگان کلیدی: هوش مصنوعی؛ کیفیت داده‌ها؛ طراحی سامانه هوشمند؛ طرح هزینه و درامد خانوار.
* نویسنده عهده‌دار مکاتبات

از پرسشگر، تلفنی، پستی، خود اجرا، از طریق اینترنت و...) اشاره کرد. بدیهی است که اطلاعات گردآوری‌شده ممکن است با نقص‌های زیادی رو به رو باشند که از جمله می‌توان به وجود داده‌های گم‌شده، دور افتاده و نادرست اشاره کرد. چنان‌چه به وجود این خطاها پی‌ببریم باید تصمیم گرفت که یا این‌گونه مشاهده‌ها را از مجموعه داده‌ها کنار گذاشت (که در پاره‌ای از موارد ممکن نیست) یا این‌که برای استفاده از این‌گونه داده‌ها باید روش‌های مناسبی را اتخاذ کرد. از طرفی باید به اندازه داده‌ها نیز به‌عنوان مقوله‌ای جدا توجه داشت. روش‌های زیادی وجود دارند که در مورد داده‌های با اندازه کم کارا هستند ولی چنان‌چه اندازه داده‌ها افزایش یابد، کارایی خود را از دست می‌دهند.

از طرفی نه تنها داشتن داده‌هایی با خطای کم در مراکز آماری اهمیت ویژه‌ای دارد بلکه عوامل دیگری نیز وجود دارند که نبود آن‌ها (حتی با وجود داشتن داده‌هایی با خطای کم) به معنی داشتن داده‌هایی با کیفیت پایین تلقی می‌شود. به‌عنوان مثال می‌توان به زمان انتشار اطلاعات آماری اشاره کرد. بدیهی است چنان‌چه بین زمان تولید و انتشار اطلاعات آماری فاصله زیادی وجود داشته باشد، انتشار آمارهایی که روزآمد نباشند، کارایی لازم را نخواهند داشت. مفهوم و عوامل مؤثر بر کیفیت داده‌ها به موارد یاد شده محدود نمی‌شود و نیاز به بحثی جداگانه دارد که در این مقاله به آن نخواهیم پرداخت. برای اطلاع بیش‌تر به [۱] و [۱۰] رجوع شود.

در بسیاری از موارد برای کاهش خطا از روش‌هایی که متکی بر نظر کارشناسان است، استفاده می‌شود. بدیهی است روش‌های ذهنی از دقت کافی برخوردار نیست و به سلیقه کارشناسان (به‌عنوان مثال در خصوص انتخاب داده دور افتاده) متکی است و به‌ویژه برای مجموعه داده‌های بزرگ، روشی دشوار و پرخطا است. امروزه روش‌ها به‌گونه‌ای طراحی می‌شوند که در حد امکان از اعمال نظر فردی کاسته شود. البته نظرهای کارشناسی صحیح، همواره در بهبود سامانه‌ها مورد استقبال قرار گرفته است.

به‌طور خلاصه می‌توان گفت به‌منظور یافتن راه‌حلی مناسب برای کاهش خطاها با چند مبحث جداگانه رو به رو هستیم: نخست، انتخاب روش مقابله با انواع خطاها، دوم مجموعه داده‌های بزرگ (به‌ویژه پایگاه داده‌ای گسترده در سازمان‌ها و مراکز آماری)، سوم کاهش اعمال نظرهای کارشناسی.

در نظر گرفتن موارد یاد شده ما را به‌سوی راه‌حلی مناسب هدایت می‌کند. یکی از راه‌های پیشنهادی استفاده از روش‌های داده‌کاوی است. در زمینه کاربردهای داده‌کاوی در آمار رسمی پژوهش‌های فراوانی صورت گرفته است. هر چند تعداد آن‌ها با توجه به اهمیت موضوع محدود است (به‌دلیل نو بودن مبحث داده‌کاوی در آمار رسمی) اما با توجه به کاربردهای داده‌کاوی، تعداد آن‌ها رو به افزایش است ([۲]، [۴]، [۵]، [۱۴] و [۱۵]).

یکی از شاخه‌های داده‌کاوی، هوش مصنوعی است که مورد نظر نویسندگان این مقاله برای پاسخ به مشکلات بیان شده است. هوش مصنوعی با توجه به قابلیت‌هایی که دارد می‌تواند ابزار مناسبی برای افزایش کیفیت داده‌ها به‌ویژه در آمار رسمی باشد [۵]، [۹]، [۱۱] و [۱۲]. با داشتن برنامه‌ها و سامانه‌های هوشمند به‌ویژه در مراکز نگهداری و گردآوری اطلاعات آماری می‌توان بر بسیاری از خطاهای غیر نمونه‌گیری فائق آمد (به‌عنوان مثال می‌توان از تکرار اشتباهات پیشین جلوگیری کرد). یکی از فواید داشتن چنین سامانه‌هایی، صرفه‌جویی در زمان (عامل مهم در قرن ارتباطات) است که مورد توجه مراکز آماری است. زیرا این مراکز باید اطلاعات آماری روزآمد و به‌هنگامی را انتشار دهند.

در بخش دوم مقاله در حد اختصار هوش مصنوعی و شاخه‌های مرتبط با آن از جمله سامانه‌های خبره و یادگیری ماشین معرفی می‌شود. در بخش سه به معرفی خطاهای مرسوم در طرح‌های آمارگیری که موجب کاهش کیفیت داده‌ها می‌شوند، پرداخته می‌شود و در بخش چهار طرح هزینه و درامد خانوار، ویژگی‌ها و چگونگی گردآوری داده‌های این طرح به‌طور خلاصه مورد بررسی قرار می‌گیرد. با استفاده از اطلاعات حاصل از بخش سه، رهنمودها و پیشنهادهایی برای طراحی یک سامانه هوشمند به‌منظور بهبود کیفیت داده‌های حاصل از این طرح در بخش پنج ارائه و ویژگی‌های آن مورد بحث قرار می‌گیرد.

۲- هوش مصنوعی

امروزه از هوش مصنوعی (Artificial Intelligence: AI) به‌عنوان دانشی رو به پیشرفت نام برده می‌شود که با توجه به قابلیت‌ها و توانایی‌هایی که دارد می‌تواند بر بسیاری از خطاها فائق آمده و باعث کاهش خطاهای غیر نمونه‌گیری در مرحله‌های گردآوری و پردازش داده‌ها شود.

کاربردهای عملی سامانه‌های هوشمند وقتی روشن‌تر می‌شوند که به‌دنبال خودکارسازی یک فرایند باشیم به‌گونه‌ای که سامانه خودکار (Automated Systems) این توانایی را داشته باشد که نه تنها فعالیت‌های روزمره و تکراری (مانند ذخیره کردن اطلاعات و محاسبات مربوط) را بلکه کارهایی که نیاز به سطحی از هوشمندی دارد، نیز انجام دهد. روشن است که طراحی یک سامانه خودکار با چنین قابلیتی مستلزم طراحی یک سامانه هوشمند در داخل فرایند خودکارسازی است که بتواند رفتار انسانی را که در فعالیت‌ها دخالت دارد (به‌عنوان مثال تصمیم گیری یا استنتاج) شبیه‌سازی کند. به‌عنوان مثال اگر فردی که در یک فرایند سازمانی حضور دارد و فردی خبره است و بر اساس تجربه‌های قبلی تصمیمی را می‌گیرد یا عملی را انجام می‌دهد، سامانه هوشمند نیز باید این قابلیت‌ها را دارا باشد؛ یعنی نخست یک سامانه، خبره باشد و دوم قدرت استفاده از تجربه‌های قبلی را دارا باشد. به عبارت دیگر سامانه قابلیت یادگیری نیز داشته باشد (برای اطلاعات بیش‌تر به [۶] مراجعه شود). بدیهی است که فرایند بهبود کیفیت داده‌ها نیز فرایندی است که تا حدی به رفتار هوشمندانه انسان و همچنین نظرهای کارشناسی افراد خبره نیاز دارد، بنا بر این برای طراحی سامانه‌ای به‌منظور بهبود کیفیت داده‌ها به استفاده از سامانه‌های خبره و یادگیری ماشین که از مهم‌ترین شاخه‌های هوش مصنوعی هستند، نیاز خواهیم داشت.

بررسی سامانه‌هایی که کل دانش مربوط به انجام یک عمل مانند تصمیم‌گیری، قضاوت یا استنتاج را داشته باشند، موضوعی از شاخه AI است که به طراحی سامانه‌های خبره مشهور است. تقریباً تمام سامانه‌های خبره قابلیت یادگیری دارند (برای اطلاعات بیش‌تر در مورد سامانه‌های خبره به [۳]، [۸] و [۱۳] مراجعه شود).

یادگیری ماشین شاخه‌ای از هوش مصنوعی است که سعی دارد الگوریتم‌هایی طراحی کند که با مشاهده و تجربه بتوانند خود را اصلاح و بهینه‌سازی کنند. در حقیقت یک سامانه خبره، هر چند پیچیده، بدیع و پیشرفته باشد، اگر نتواند یاد بگیرد تا اشتباه‌های گذشته را تکرار نکند، کارا نیست و در عمل بدون استفاده خواهد ماند (برای اطلاع بیش‌تر به [۱۷] و [۱۸] مراجعه شود).

۳- خطاهای آمارگیری

به‌طور کلی، خطاهای آمارگیری را می‌توان به دو رده اصلی خطاهای نمونه‌گیری و خطاهای غیر نمونه‌گیری، تقسیم کرد. خطاهای نمونه‌گیری، از انتخاب آگاهانه بررسی بخشی از جامعه به‌جای کل جامعه، ناشی می‌شود و می‌توان با به‌کارگیری روش‌های مناسب برآورد و انتخاب تصادفی، برای کنترل آن تلاش لازم را به عمل آورد. خطاهای نمونه‌گیری، اشتباه به حساب نمی‌آیند ولی می‌توانند هنگام استنباط موجب بروز اشتباه‌های بزرگ شوند.

به غیر از خطاهای نمونه‌گیری، سایر خطاها را خطاهای غیر نمونه‌گیری می‌نامند. در بیش‌تر موارد، خطاهای غیر نمونه‌گیری به اشتباه‌هایی گفته می‌شود که در مراحل مختلف تهیه و اجرای یک طرح آمارگیری اتفاق افتاده و موجب کاهش سودبخشی و مطلوبیت نتایج آن می‌شوند. این‌گونه خطاها به سبب دلایلی مانند تعریف‌های نادرست، برنامه‌های گزارش‌گیری نامناسب، عدم موفقیت در کسب اطلاعات لازم از تمام واحدهای نمونه، خطاهای ناشی از اعمال سلیقه‌های شخصی و... بروز می‌کنند.

بنا بر این برای کنترل مجموع خطاهای برآورد حاصل از تمام منابع مختلف خطا در طراحی آمارگیری تلاش‌هایی صورت می‌گیرد. معمولاً آمارگیری‌ها با در نظر گرفتن تمامی منابع شناخته‌شده خطا، به‌دقت برنامه‌ریزی‌شده و تمامی منابع در دسترس آمارگیری به‌منظور کمینه کردن مجموع خطا مورد استفاده قرار می‌گیرند. به موازات این اقدام، پژوهشگران با استفاده از روش‌های کنترل کیفی، به هنگام وقوع اشتباه‌ها در مرحله گردآوری داده‌ها و فراوری داده‌ها هشدارهای لازم را می‌دهند. در مرحله تحلیل نیز کوشش می‌شود حدود احتمالی کل خطای برآورد، تعیین و در گزارش‌های مربوط آورده شود. همچنین برآورد مؤلفه‌های کل خطا به‌گونه‌ای تعیین می‌شوند که در برنامه‌ریزی‌های مربوط به آمارگیری‌های بعدی نیز قابل استفاده باشند. در مسیر تلاش برای پیاده‌سازی طرح آمارگیری در مراحل طراحی، اجرا یا تحلیل آمارگیری‌ها، مشکلات و پیچیدگی‌هایی وجود دارند که تا حد ممکن باید برای رفع آن‌ها سعی کرد. به‌عنوان مثال، امکان دارد اطلاعات کمکی دیگری برای طرح لازم باشد. بنا بر این باید روش‌های تجربی خاصی به کار گرفته شوند تا بتوان اثر منبع معینی از خطای غیر نمونه‌گیری را برآورد کرد. از طرفی

چنان‌چه اطلاعات لازم در دسترس است، منابع مالی نیز باید در نظر گرفته شوند. همچنین روش‌های کنترل کیفی نیز هزینه‌بر هستند و این موضوع نیز باید مدنظر قرار گیرد. بنا بر این توصیه می‌شود از سامانه‌های هوشمند در این رابطه استفاده شود [۷].

۴- طرح هزینه و درامد خانوار و مراحل آن در یک نگاه

در ادامه به کاربردی از هوش مصنوعی در آمار رسمی می‌پردازیم. به این منظور از طرح هزینه و درآمد خانوار که هر سال توسط مرکز آمار ایران انجام می‌شود، استفاده شده است. شایان ذکر است رهنمودها و پیشنهادهایی که در این نوشتار برای طراحی یک سامانه هوشمند در مورد طرح یاد شده ارائه خواهد شد به سایر طرح ها نیز قابل تعمیم است.

۴-۱- مراحل اجرایی فعالیت‌های مرتبط با کیفیت داده‌های آماری در طرح هزینه و درامد خانوار

معمولاً در فرایند گردآوری داده‌ها فعالیت‌هایی برای بررسی تضمین کیفیت داده‌های گردآوری‌شده صورت می‌گیرد. در حال حاضر در طرح هزینه و درامد خانوار نیز چنین فعالیت‌هایی در حال انجام است. پیش از آن‌که به کلیات یک سامانه هوشمند برای بهبود کیفیت داده‌ها در این طرح پرداخته شود نگاهی کوتاه بر مراحل اجرایی فعالیت‌های مرتبط با کیفیت داده‌ها در این طرح انداخته شده و برخی از آن‌ها بررسی شده است:

آ) اولین گام برای تضمین کیفیت داده‌ها سعی در گردآوری داده‌های صحیح از منبع گردآوری داده‌ها است. به همین دلیل قبل از اجرای طرح در استان، آمارگیرها آموزش‌های لازم را برای ثبت داده‌های صحیح و مطمئن دریافت می‌کنند؛

ب) در گام بعدی، هنگامی که داده‌ها گردآوری شدند، در هر استان پرسشنامه‌ها در اختیار بازبین‌ها قرار می‌گیرند تا بر اساس دستورالعمل بازبینی، کار ویرایش اولیه پرسشنامه را برای تصحیح احتمالی انجام دهند و کیفیت داده‌های ثبت‌شده را مورد بررسی قرار دهند. (البته این فرایند خود ممکن است تولید خطا کند، چرا که به تجربه و دقت بازبین بستگی دارد). اگر پرسشنامه‌ای به تشخیص بازبین، حاوی داده‌های نادرست

باشد به آمارگیر برگردانده می‌شود تا اشتباه‌های پرسشنامه را تصحیح کند؛

دستورالعمل بازبینی دارای دو دسته از قاعده‌ها است که در ادامه شرح داده شده‌اند، اگر هر یک از آن‌ها در پرسشنامه رعایت نشده باشد یا باید علت عدم رعایت آن توسط آمارگیر توضیح داده شود یا پرسشنامه برای اصلاح در اختیار آمارگیر قرار گیرد. دسته اول قاعده‌هایی صریح‌اند (به‌عنوان مثال جمع برخی از سطرها باید با سطر مشخصی برابر باشد). بدیهی است بررسی چنین قاعده‌هایی نیاز به حضور یک موجود هوشمند ندارد. دسته دوم قاعده‌هایی‌اند که به اندازه دسته اول صریح نیستند (به‌عنوان مثال هزینه‌های مربوط به تهیه برخی مواد باید معقول باشد). بررسی چنین قاعده‌هایی به درجه‌ای از هوشمندی، خبرگی و همچنین قابلیت یادگیری نیاز دارد.

پ) بعد از بررسی اولیه داده‌ها، وارد کردن داده‌ها صورت می‌گیرد، با توجه به پرسشنامه گسترده این طرح پس از وارد کردن داده‌ها، تصدیق یا وریف[1] صد در صد روی داده‌ها انجام می‌شود تا از سلامت داده‌هایی که به نرم‌افزار واردشده اطمینان حاصل شود. (بدیهی است اگر بتوان به‌گونه‌ای از سلامت داده‌ها هنگام ورود به پایگاه داده‌ها اطمینان حاصل کرد دیگر نیازی به تصدیق صد در صد داده‌ها نخواهد بود)؛

ت) کارشناس مسئول طرح در استان در دو مرحله، چگونگی تکمیل پرسشنامه را مورد ارزیابی قرار می‌دهد. در مرحله اول پیش از وارد کردن داده‌ها و در مرحله دوم پس از تصدیق و ویرایش آدرس[2]. در این مرحله کارشناس به ویرایش موضوعی[3] پرسشنامه‌ها با استفاده از نرم‌افزار ویرایش موضوعی[4] می‌پردازد. در این مرحله برخی از خطاهای قابل شناسایی که در بازبینی ممکن است دور از نظر مانده باشند و همچنین خطاهای وارد کردن داده‌ها و تصدیق داده‌ها مشخص و برطرف می‌شود؛

ث) پس از ویرایش موضوعی اولیه در استان، پرونده اطلاعات به مرکز آمار ایران ارسال می‌شود. در این مرحله خطاهای احتمالی که در استان‌ها قابل شناسایی و رفع نبوده است، توسط کارشناس موضوعی مورد بررسی قرار گرفته و در صورت امکان با هماهنگی استان برطرف می‌شوند؛

ج) علاوه بر مراحل بالا، به‌منظور کنترل خطای نمونه‌گیری آمارشناس دفتر استانداردهای

آماری (یکی از دفترهای مرکز آمار ایران) بـا اسـتفاده از فـرض‌هـایی کـه کارشـناس موضوعی در اختیارش قرار داده و راهنمای تشخیص، اندازه نمونه را تعیین می‌کنـد تـا بیش‌ترین دقت در تحلیل‌های بعدی حاصل شود. در طول اجرای طـرح ممکـن اسـت خانواری در دسترس نباشد یا بنا به‌دلیلی امکان تکمیل پرسشنامه را نداشـته باشـد، در این صورت خانوار دیگری جانشین آن می‌شود. بنا بر این اگر در طول اجرای طرح نیاز به جایگزین کردن نمونه‌ها باشد ایـن اقـدام بـا همـاهنگی آمارشـناس طـرح صـورت می‌گیرد.

۲-۴- برخی از نکات برجسته در تحلیل این فرایند

در بررسی این فرایند چند نکته اساسی وجود دارند:

آ) در این سامانه، داده‌های هر پرسشنامه (به‌جز تصدیق) چهار بار و توسط سه فرد متفاوت بررسی می‌شوند. بازبین در استان، کارشناس طرح در استان پیش از وارد کردن داده‌هـا، کارشناس طرح در استان پس از تصدیق داده‌ها و درنهایت ویـرایش موضـوعی توسـط کارشناس موضوعی در مرکز آمار ایران؛

ب) در تمام مراحل بررسی، این احتمال وجود دارد که پرسشنامه برای تکمیل بـه آمـارگیر برگردانده شود تا دوباره به خانوار مراجعه کند. ایـن مـسئله مـی‌توانـد تـا حـد زیـادی وقت‌گیر و هزینه‌بر باشد؛

پ) چون احتمال خطای انسانی در کارهـای تکـراری ماننـد وارد کـردن حجـم زیـادی از اطلاعات یا محاسبات پیچیده و تکراری، به نسبت زیاد است، بازبینی دستی در مـورد قاعده‌هایی که صریح‌اند از یک‌سو و همچنین وارد کردن داده‌ها توسط انسان از سوی دیگر، می‌توانند باعث وقوع خطاهـای غیـر نمونـه‌گیـری جدیـدی در داده‌هـا شـوند. همچنین انجام چنین فعالیت‌هایی، توسط انسان، به‌دلیل کاهش کارایی عامل انسانی، باعث پایین آمدن بازدهی فرایند می‌شود؛

ت) با انجام همه این اعمال هنوز همه مؤلفه‌های کیفیت داده‌ها مورد بررسی قرار نگرفتـه است. در سامانه بالا اطلاعات هر خانوار بدون توجه به جامعه‌ای کـه در آن قـرار دارد

مورد بررسی قرار گرفته است. حال آن‌که ممکن است داده‌های مربوط به یک خانوار با توجه به دستورالعمل‌های بازبینی قابل قبول باشد ولی در جامعه‌ای که خانوار در آن قرار دارد چنین داده‌هایی چندان محتمل نباشند. بررسی این بی‌نظمی‌های پنهان، مستلزم استفاده از یک روش هوشمند است. از آن‌جایی که کاوش در حجم بالای داده‌ها نیازمند محاسبات سنگین و درجه‌ای از هوشمندی است، بررسی این جنبه از کیفیت داده‌ها بدون کمک نرم‌افزارهای داده‌کاوی می‌تواند خود منبع وقوع خطای غیر نمونه‌گیری جدید باشد.

با توجه به آن‌چه گفته شد، سامانه هوشمندی که برای بهینه‌سازی فرایند بهبود کیفیت داده‌ها در طرح هزینه و درامد خانوار، طراحی می‌شود باید توانایی رفع نقص‌های سامانه موجود را دارا باشد. بنا بر این:

آ) این سامانه هوشمند باید بتواند بر اساس دستورالعمل بازبینی در مورد درستی داده‌های یک پرسشنامه قضاوت کند. به‌ویژه باید بتواند قاعده‌های غیر صریح دستورالعمل را به کار گیرد. از آن‌جایی که قضاوت بر اساس این قاعده‌ها توسط انسان خبره صورت می‌گیرد، سامانه هوشمند نیز باید توانایی سامانه خبره را دارا باشد. همچنین چون در میان قاعده‌های غیر صریح در برخی موارد قاعده‌های نامطمئن نیز وجود دارند، سامانه ممکن است در شرایطی به اشتباه قضاوت کند؛ پس علاوه بر خبرگی باید دارای توانایی یادگیری نیز باشد (تا اشتباه‌های پیشین را تکرار نکند)؛

ب) این سامانه هوشمند باید توانایی یافتن بی‌نظمی‌های پنهان در داده‌ها را (با کاوش در داده‌های گردآوری‌شده) داشته باشد.

۵- پیشنهادهایی برای طراحی یک سامانه هوشمند

برای طراحی یک سامانه هوشمند (Data Control Intelligent System 1: DCIS1) که توانایی‌های یک گروه از عامل‌های هوشمند را دارا باشد، قبل از هر چیز باید آگاهی صحیحی از رفتار گروهی این عامل‌ها داشت. بنا بر این گام نخست در طراحی یک سامانه هوشمند، تحلیل رفتار عامل‌های هوشمند به‌منظور ارائه مدلی منطبق با روند تصمیم‌گیری

گروهی آن‌ها است. به این معنی که سامانه هوشمند باید وظیفه محول‌شده به سامانه مورد نظر را انجام دهد.

در گام بعدی نکات برجسته‌ای مشخص شده‌اند که در حقیقت ویژگی‌های اصلی سامانه هوشمندند و باید در طراحی سامانه مورد توجه قرار گیرند. در این گام جزئیات مسئله‌ای که باید حل شود دقیقاً مشخص می‌شوند.

پس از مشخص شدن مأموریت‌های سامانه هوشمند، با استفاده از نتایجی که از مطالعه رفتار گروهی و روش‌هایی که برای طراحی یک سامانه هوشمند یا خبره به کار می‌روند، ابتدا قالب سامانه هوشمند و سپس جزئیات بیش‌تر طراحی و اجرا می‌شوند.

پس از انجام این مراحل، تمام فعالیت‌ها روی بهینه‌سازی و تکمیل سامانه طراحی‌شده، متمرکز خواهند شد تا فرایند مهندسی دانش سامانه تکمیل شود و سامانه به‌طور کامل به خود متکی شود.

برای کاهش پیچیدگی الگوریتم‌های یک سامانه هوشمند و تسهیل طراحی و بهینه‌سازی سامانه، عموماً از معماری چند لایه در طراحی چنین سامانه‌هایی استفاده می‌شود. این عمل باعث تسهیل در اجرای سامانه طراحی‌شده خواهد شد. همچنین استفاده از معماری چند لایه قابلیت حل مسئله به کمک راه‌حل‌های موجود را دارا خواهد بود. به این معنی که به‌جای حل مسئله از ابتدا تا مراحل نهایی و تعریف تمام جزئیات سامانه، این قابلیت وجود خواهد داشت که یک مسئله به اجزای کوچک‌تری تقسیم شود که برای این اجزاء کوچک‌تر راه‌حل‌هایی موجودند. به این ترتیب برای طراحی سامانه هوشمند کافی است اجزای مختلفی که مسائل مشخصی را حل می‌کنند، به‌گونه‌ای کنار یکدیگر (در یک معماری چند لایه) چیده شوند که رفتاری هوشمندانه از خود بروز دهند [۱۶].

۱-۵- معرفی قالب DCIS1

با توجه به ویژگی‌هایی که پیش‌تر برای سامانه هوشمند کنترل داده‌ها از آن‌ها یاد شد، معماری چنین سامانه‌ای می‌تواند به شکل زیر باشد (برای درک بهتر معماری سامانه، سامانه‌ای چند لایه پیشنهاد می‌شود).

لایه اول: این لایه شامل همه الگوریتم‌ها و فنونی است که در بازبینی و ویرایش موضوعی مورد استفاده قرار می‌گیرند. این لایه بر اساس دستورالعمل بازبینی و نمودار ویرایش موضوعی، داده‌های مربوط به هر خانوار را بررسی کرده و خطاهای احتمالی موجود در مجموعه داده‌ها را می‌یابد. به عبارت دیگر خروجی این لایه، صفت‌هایی از هر خانوارند که از نظر سامانه ممکن است صحیح نباشند.

لایه دوم: این لایه خروجی لایه پیشین را دریافت کرده و آن‌ها را بررسی می‌کند. بر اساس این‌که سامانه تا چه میزانی در مورد نادرست بودن هر یک از صفت‌های گزارش‌شده از لایه اول اطمینان دارد، تصمیم‌هایی گرفته می‌شوند و بر اساس این تصمیم‌ها داده‌ها اصلاح می‌شوند.

پس از اعمال لایه دوم ممکن است پرسشنامه مربوط به بعضی از خانوارها برای اصلاح به آمارگیر برگردانده شود. داده‌هایی که به این شیوه اصلاح می‌شوند بعد از گردآوری، یک بار دیگر وارد لایه اول می‌شوند (لایه اول لایه ورودی سامانه است، بنا بر این هر داده‌ای که بخواهد وارد سامانه شود باید از این لایه عبور کند).

لایه سوم: در دو لایه پیشین همه خطاهایی که در سامانه فعلی قابل شناسایی‌اند، شناسایی شده و رفع شده‌اند. این لایه (برای اطمینان از صحت عملکرد دو لایه پیشین) با استفاده از فنونی ساده به بررسی ساختار اطلاعاتی نمونه می‌پردازد. عمده‌ترین روش مورد استفاده در این لایه رسم جدول‌ها با استفاده از روش پردازش تحلیلی بر خط (On Line Analytical Process: OLAP) و در کنار آن استفاده از روش‌های توصیفی برای تبیین ساختار نمونه موجود است. بخشی از خطاهایی که در دو لایه پیشین ممکن است پنهان مانده باشند، با تحلیل جدول‌های به دست آمده از روش پردازش تحلیلی بر خط و خروجی‌های توصیفی داده‌ها نمایان می‌شوند. بدیهی است اگر خطایی در ثبت داده‌ها وجود داشته باشد جدول‌های OLAP و خروجی‌های توصیفی داده‌ها، ساختاری غیر منطقی برای داده‌ها ترسیم خواهند کرد.

به‌طور خلاصه، لایه سوم با استفاده از روش‌های آمار توصیفی و جدول‌های به دست آمده از OLAP، بخش‌هایی از داده‌ها را که دارای ساختاری غیر منطقی هستند شناسایی

کرده و آن‌ها را به‌عنوان خروجی گزارش می‌کند.

لایه چهارم: این لایه با استفاده از اطلاعاتی که از لایه سوم (خطاهای گزارش‌شده در ساختار داده‌ها) در مورد خطاهای گزارش‌شده به دست می‌آید، تصمیم‌گیری می‌کند. بخشی از این خطاها ممکن است در درون همین لایه قابل اصلاح باشند که در این صورت سامانه در همین لایه آن‌ها را اصلاح می‌کند. برای اصلاح بخشی از این خطاها که توسط این لایه قابل برطرف شدن نیستند باید پرسشنامه‌ها به آمارگیر بازگردانده شوند. خروجی این لایه علاوه بر داده‌های اصلاح‌شده مواردی همچون فراداده‌ها و اطلاعاتی درباره ساختار داده‌ها را نیز شامل می‌شود. شایان ذکر است که داده‌های اصلاح‌شده توسط آمارگیر، خود ممکن است اشتباه باشند و سامانه نیز نتواند آن‌ها را شناسایی کند بنا بر این خطا همچنان در داده‌ها باقی می‌ماند.

لایه پنجم: لایه پنجم به یافتن بی‌نظمی‌های پنهان موجود در داده‌ها می‌پردازد. در حقیقت این لایه یک برنامه داده‌کاوی است که در پایگاه داده‌ها به‌دنبال خطاهایی است که در مراحل پیشین پنهان مانده‌اند. این لایه با بررسی موقعیت بردار داده‌های هر خانوار در فضای داده‌ای موجود سعی در یافتن خانوارهایی دارد که به هر دلیل در بخش خاصی از فضای داده‌ای به‌صورت انبوهه در آمده‌اند. برای رسیدن به این هدف، این لایه از سامانه به خوشه‌بندی خانوارها می‌پردازد (روش‌های خوشه‌بندی در این قسمت بسیار کارا هستند). خروجی این لایه مقادیر مربوط به خانوارهایی هستند که در بخشی از فضای داده‌ای به‌صورت انبوهه در آمده‌اند (به عبارتی دیگر در یک خوشه قرار گرفته‌اند). به این منظور در نظر گرفتن صفت‌های مشترک خانوارهایی که در هر موقعیت از فضای داده‌ای قرار گرفته‌اند بسیار مفید و آگاهی‌بخش است. شایان ذکر است در طراحی این لایه نیز می‌توان از یک ساختار چند لایه‌ای دیگر استفاده کرد. همچنین به‌نظر می‌رسد به‌جای طراحی این لایه از پایه، می‌توان الگوریتم‌های داده‌کاوی موجود را برای استفاده در این لایه بهینه کرد.

لایه ششم: این لایه با دریافت خروجی لایه پنجم به بررسی دلایل انبوهش بردار خانوارها در هر بخش از فضای داده‌ای می‌پردازد. این بررسی با استفاده از اطلاعاتی که از

لایه پنجم به دست می‌آید، امکان‌پذیر است. به‌عنوان مثال می‌توان به بررسی تعداد خانوارهایی که در یک خوشه انبوهش یافته‌اند، صفت‌های مشترک این خانوارها، تعداد و نوع صفت‌ها اشاره کرد. با یافتن دلایل انبوهش داده‌ها، می‌توان داده‌ها را به دو دسته تقسیم کرد. دسته اول داده‌هایی‌اند که بنا بر دلایل منطقی در یک بخش ویژه از فضای داده‌ای انبوهش یافته‌اند و دسته دوم که هیچ دلیل منطقی برای انبوهش آن‌ها وجود ندارد. دسته دوم را می‌توان داده‌هایی در نظر گرفت که دلیل انبوهش آن‌ها وقوع یک خطای مشترک بوده است. در فرایند بررسی دلایل انبوهش داده‌ها منابع خطاها و خطاهای مشترک نیز به دست خواهند آمد.

این لایه می‌تواند در مورد دسته‌ای که برای انبوهش آن‌ها دلیلی منطقی وجود دارد، به تولید فراداده‌هایی در مورد این بخش از داده‌ها بپردازد تا از این پس تحلیل‌ها روی مجموعه داده‌ها با توجه به این فراداده‌ها انجام شود. خروجی این لایه شامل دسته‌بندی خوشه‌های ساخته‌شده در لایه پنجم و ارائه گزارش درباره دلایل تشکیل خوشه‌ها است. در مورد خوشه‌هایی که وقوع خطایی در گردآوری داده‌ها را نمایش می‌دهند، این گزارش شامل منابع وقوع خطا و همچنین نقص‌هایی در داده‌ها است که به‌دلیل وقوع این خطا به وجود آمده‌اند. این گزارش، فراداده‌های تولیدشده برای هر خوشه را در مورد خوشه‌هایی که دلیلی موجه برای تشکیل داشته‌اند نیز شامل می‌شود.

لایه هفتم: این لایه، خوشه‌هایی را که بر اثر وقوع خطا تشکیل شده‌اند، اصلاح می‌کند. این اصلاح با توجه به گزارشی که از لایه پیشین به دست آمده است، انجام می‌شود. در این مرحله ممکن است تعدادی از پرسشنامه‌ها برای اصلاح به آمارگیر برگردانده شوند. هر چند که هنوز احتمال این‌که به اشتباه توسط آمارگیر اصلاح شوند وجود دارد.

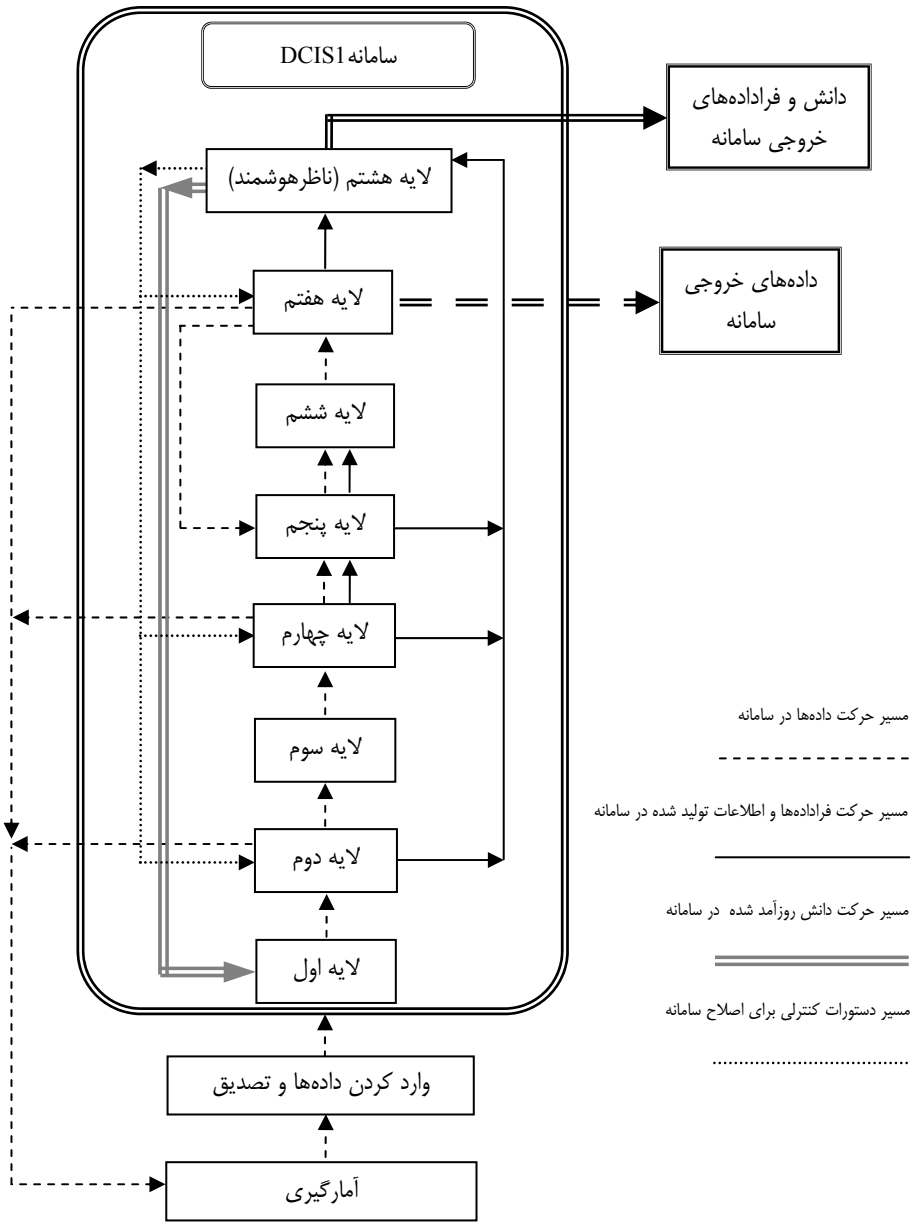
لایه هشتم (سامانه نظارت و اصلاح): در لایه‌های مختلفی که تا این‌جا مورد بحث قرار گرفته‌اند، فراداده‌ها و اطلاعات جدیدی تولید شده‌اند. لایه هشتم با استفاده از اطلاعات و فراداده‌هایی که در طول فرایند تولید شده‌اند، بر نحوه انجام فرایند نظارت کرده و در بخش‌هایی آن را اصلاح می‌کند. مهم‌ترین بخش در این لایه استفاده از فراداده‌ها و اطلاعات برای اصلاح سامانه است. برای این‌که بتوانیم سامانه را به‌گونه‌ای کارامد اصلاح

کرده و به‌روز نگاه داریم، باید بتوانیم از اطلاعات گردآوری‌شده، تولید دانش و فرادانش کنیم. مهم‌ترین اطلاعاتی که می‌تواند به کم شدن احتمال عدم شناسایی یک داده اشتباه یا دورافتاده کمک کند، اطلاعات و فراداده‌هایی است که در لایه پنجم تولید می‌شوند. این اطلاعات شامل دلایل وقوع خطا، نقص‌هایی که در داده‌ها به‌دلیل وقوع خطا رخ می‌دهند و دلایل انبوهش منطقی داده‌ها در یک موقعیت ویژه است. با استفاده از این اطلاعات در این لایه، دانش گذشته برای بازبینی پرسشنامه‌ها به‌روز می‌شود. علاوه بر این داده‌ها و اطلاعات دیگر نظیر تعداد خطاهای پذیرفته شده در سامانه، خود سامانه نیز این امکان را ایجاد می‌کند تا بر عامل‌های مختلف در سامانه نظارت شود و خروج هر عامل از حالت بهینه به این وسیله اصلاح گردد. بنا بر این لایه هشتم دو وظیفه مهم دارد:

اول: آن‌که تغییرهای درون سامانه را مورد بررسی قرار دهد تا از خارج شدن سامانه از حالت بهینه جلوگیری کند (کنترل فرایند).

دوم: آن‌که با استفاده از فراداده‌ها و اطلاعاتی که در طول فرایند تولید شده، به تولید دانش بپردازد و دانش تولیدشده را به لایه‌های پایین‌تر منتقل کند.

شکل زیر نمای کلی سامانه DCIS1 را ارائه می‌کند.



شکل ۱- نمای کلی سامانه DCIS1

۲-۵- نکاتی درباره سامانه DCIS1

درباره سامانه هوشمندی که کلیات آن را ارائه کردیم چند نکته مهم وجود دارند:

آ) در این سامانه بعد از عبور از لایه پنجم هر بار که اصلاحی روی بخشی از داده‌ها انجام شود، همه داده‌ها (حتی داده‌هایی که هیچ اصلاحی بر روی آن‌ها انجام نشده) باید دوباره به لایه پنجم بازگردانده شوند. این عمل به آن دلیل است که لایه پنجم موقعیت داده‌ها را نسبت به هم بررسی کرده و بر این اساس خروجی تولید می‌کند. بنا بر این هر تغییری در داده‌ها ممکن است باعث تغییر در موقعیت داده‌ها نسبت به یکدیگر شود. به عبارت دیگر هر بار که قرار است داده‌ای در لایه پنجم بررسی شود، همه داده‌ها باید به این لایه وارد شوند؛

ب) این سامانه یک سازماندهی هوشمند است که با توزیع دانش در میان اجزای مختلف، یک هوش توزیع‌یافته را در کل فرایند به اجرا درمی‌آورد. به عبارت دیگر، DCIS1 دارای یک هوش جمعی مبتنی بر توزیع دانش و هم‌جوشی اطلاعات است. به این معنی که حیاتی‌ترین ویژگی این سامانه، فرایند گردش اطلاعات و توزیع دانش است. به همین دلیل هرچه توزیع دانش و گردش اطلاعات کامل‌تر انجام گیرد، سامانه رفتار هوشمندانه‌تری از خود بروز می‌دهد؛

پ) پویایی قالب DCIS1، این اجازه را به سامانه می‌دهد که در یک سازمان متشکل از عامل‌های انسانی یا عامل‌های شبیه‌سازی‌شده به اجرا درآید. اما با توجه به ساختار درونی سامانه، بهترین سازمان برای به اجرا درآوردن سامانه، سازمانی متشکل از عامل‌های انسانی (بیش‌تر به‌عنوان ناظر) و عامل‌های شبیه‌سازی‌شده (به‌عنوان تصمیم‌گیرنده یا پشتیبان تصمیم‌گیری) است؛

ت) این سامانه در صورتی که بر پایه عامل‌های شبیه‌سازی‌شده به اجرا درآید تا حد زیادی می‌تواند در زمان و هزینه صرفه‌جویی کند. چرا که در این صورت علاوه بر استفاده از مزیت‌های خودکارسازی، بدون این‌که کیفیت را از دست بدهد تعداد بررسی‌های یک پرسشنامه را کاهش می‌دهد. (همه بررسی‌های چهارگانه که در حال حاضر در سامانه فعلی انجام می‌گیرند در لایه‌های اول و دوم انجام می‌شوند). بدیهی است هزینه و

زمانی که به این وسیله صرفه‌جویی می‌شود برای اعمال لایه‌های بالاتر روی داده‌ها می‌تواند بسیار مفید باشد؛

ث) از آن‌جا که بهترین گزینه برای به اجرا درآوردن سامانه در لایه اول، استفاده از نرم‌افزار و عامل‌های شبیه‌سازی‌شده است، لازم است اطلاعات از همان ابتدا به نرم‌افزار منتقل شوند. با توجه به این‌که وارد کردن داده‌ها با استفاده از صفحه کلید (به‌صورت دستی) خود می‌تواند باعث وقوع خطاهایی در داده‌ها شود (هر چند داده‌ها پس از وارد شدن ویرایش می‌شوند) استفاده از نرم‌افزارهای وارد کردن داده‌ها می‌تواند کارایی سامانه را تا حد زیادی افزایش دهد. ضمن این‌که استفاده از نرم‌افزار وارد کردن داده‌ها می‌تواند نیاز به تصدیق را از بین ببرد (این مسئله تا حد زیادی بستگی به‌دقت نرم‌افزار وارد کردن داده‌ها دارد). یکی از راه‌های پیشنهادی برای وارد کردن داده‌ها به نرم‌افزار، ثبت داده در رایانه در مرحله آمارگیری است (استفاده از روش رایانه‌یار)؛

ج) در نگاه اول به‌نظر می‌رسد که DCIS1 چندان قابل اجرا نبوده و با یک سامانه قابل اجرا فاصله زیادی داشته باشد. اما واقعیت این است که DCIS1 یک ساختار هوشمند ارائه می‌کند که حتی با ابزارهای فعلی نیز می‌تواند به‌خوبی رفتاری هوشمندانه بروز دهد. اگرچه در حالت ایده‌آل لایه هشتم باید از پایه طراحی شود، بقیه لایه‌ها نیازی به طراحی از پایه را ندارند. زیرا الگوریتم‌های کلی مربوط به این لایه‌ها از پیش وجود داشته و برای اجرایی ساختن آن‌ها کافی است با یک بازنگری در این الگوریتم‌ها، آن‌ها را با DCIS1 سازگار کرد. به بیان دیگر DCIS1 از ابزارهایی استفاده می‌کند که از پیش موجودند (به‌جز لایه هشتم که به‌نظر می‌رسد باید ساختار پیچیده‌تری داشته باشد). بنا بر این در اجرایی کردن DCIS1 کافی است بتوانیم از این ابزارها استفاده کنیم.

سپاس‌گزاری

نویسندگان این مقاله از مرکز آمار ایران به خاطر حمایت مالی در انجام طرح مطالعاتی داده‌کاوی و کاربرد آن در کیفیت داده‌ها، آقای عبدالحمید حقیقی به خاطر راهنمایی‌های ایشان و همچنین دفتر استانداردهای آماری سپاس‌گزارند.

توضیحات

[۱] تصدیق یا وریف: وارد کردن دوباره داده‌ها و مقایسه آن‌ها با هم برای این‌که خطاهای احتمالی وارد کردن داده‌ها بررسی شود.

[۲] ویرایش آدرس: بررسی اطلاعات مربوط به آدرس و چارچوب.

[۳] ویرایش موضوعی: بررسی رابطه بین فیلدهای پرسشنامه و جلوگیری از ورود اطلاعات غلط داخل فایل (بررسی رابطه منطقی بین پرسش‌ها).

[۴] نرم‌افزار ادیت موضوعی: توسط تحلیل‌گر سامانه و بر اساس نمودار ویرایش موضوعی، که توسط کارشناس موضوعی طراحی شده است، تهیه می‌شود.

مرجع‌ها

[۱] اناری، محمدرضا (۱۳۸۴). ابعاد و ملاک‌های کیفیت و روش‌های ارزیابی کیفیت آمارهای رسمی. گزیده‌مطالب آماری، سال ۱۶، شماره ۱، صص ۳۶-۵۸.

[۲] حسنی، حسین؛ حائری مهریزی، علی‌اصغر (۱۳۸۵). داده‌کاوی و کاربرد آن در آمار رسمی. گزیده‌مطالب آماری، سال ۱۶، شماره ۴، صص ۲۱-۳۴.

[۳] دارلینگتون، کیس (۱۹۹۹). سامانه‌های خبره. همایون مؤتمنی (مترجم)، ترجمه از نسخه انگلیسی، علوم رایانه، تهران.

[۴] دفتر تعاریف و استانداردهای آماری (۱۳۸۴). گزارش اول طرح مطالعاتی داده‌کاوی و کاربرد آن در کیفیت داده‌ها. مرکز آمار ایران، تهران.

[۵] دفتر تعاریف و استانداردهای آماری (۱۳۸۵). گزارش دوم طرح مطالعاتی داده‌کاوی و کاربرد آن در کیفیت داده‌ها. مرکز آمار ایران، تهران.

[۶] راسل، استوارت؛ نورویگ، پیتر (۲۰۰۱). هوش مصنوعی رهیافتی نوین، رامین رهنمون و آناهیتا هماوندی (مترجمان)، ترجمه از نسخه انگلیسی، ناقوس، تهران.

[۷] لسلر، جودیث تی؛ کالس بک، ویلیام دی (۱۹۹۲). خطاهای غیر نمونه‌گیری در آمارگیری‌ها، مرکز آمار ایران (مترجم). ترجمه از نسخه انگلیسی، جلد اول، مرکز آمار ایران، تهران.

[۸] یگانگی، محمدرضا (۱۳۸۳). تجزیه و تحلیل الگوریتم‌های کامپیوتری از نگاه آمار، دومین سمینار مهندسی کامپیوتر، مؤسسه آموزش عالی جهاد دانشگاهی، تهران.

[۹] یگانگی، محمدرضا؛ حسنی، حسین؛ حائری مهریزی، علی‌اصغر (۱۳۸۵). هوش مصنوعی و کاربرد آن در آمار رسمی. هشتمین کنفرانس بین‌المللی آمار ایران، شیراز، صص ۱۳۴-۱۴۴.

[10] Elvers, E.; Rosen, B. (1997). *Quality concepts for official statistics, Encyclopedia of Statistical Sciences*, 621-629. Wiley. New York.

[11] Domingo-Ferrer, J.; Torra, V. (2003). On the Connections between Statistical Disclosure Control for Micro data and Some Artificial Intelligence Tools, *Inform. Sci.* **515**, 153-170.
[12] Domingo-Ferrer, J.; Torra, V. (2002). *Disclosure control methods and information loss for micro data in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Amsterdam, North-Holland, 93-112.
[13] Durkin, J. (1994). *Expert systems design and development*, Prentice-Hall.
[14] Hassani, H.; Anari, M. (2005). Using Data Mining for Data Quality Improvement, *Proceedings of the 55th session of the International Statistical Institute (ISI)*. Australia.
[15] Hassani, H.; Haeri Mehrizi A.A. (2006). Data Mining & Official Data. *Proceeding of the 8th Iranian International Statistics Conference*, 111-117. Iran.
[16] Leiserson, Charles E, et al. (2001). *Introduction to Algorithms*, MIT press.
[17] Mitchell, T. (1997). *Machine Learning*, McGraw Hill, New York.
[18] Nilsson, J. (1997). *Introduction to machine learning*, Stanford University, Stanford.