مقایسه دقت الگوریتمهای یادگیری ماشین در تخمین داده‌های گمشده حاصل از آزمایشهای ریزآرایه DNA

مریم مشیری[۱]، مصطفی قادری زفره‌ایی[۲]* و فرزان قانع گل‌محمدی[۳]

[۱] مشهد، دانشگاه فردوسی مشهد، دانشکده کشاورزی، گروه علوم دامی

[۲] یاسوج، دانشگاه یاسوج، دانشکده کشاورزی، گروه علوم دامی

[۳] کرج، پژوهشکده بیوتکنولوژی کشاورزی، گروه زیست شناسی سیستمها

تاریخ دریافت: ۹۳/۱۱/۱۲ تاریخ پذیرش: ۹۴/۷/۱۲

چکیده

وجود داده‌های گمشده در داده‌های ریزآرایه، سبب کاهش دقت رسم شبکه‌های تنظیمی ژن، ایجاد اشتباه در خوشه بندی و تقسیم‌بندی تخصصی ژنها و سایر تحلیلها می‌شود. بنابراین تخمین داده‌های گمشده مرحله مهمی در پیش پردازش داده‌های ریزآرایه، محسوب می‌شود. عملکرد الگوریتمهای تخمین در مجموعه داده‌های مختلف و با درصدهای متفاوت گمشدگی، متغیر است. همواره انتخاب مناسب‌ترین الگوریتم به منظور دستیابی به بیشترین دقت در محاسبات داده‌های گمشده از اهمیت خاصی برخوردار است. در این مطالعه از سه مجموعه داده آزمایشهای ریزآرایه استفاده شد. پس از مشخص کردن ابعاد ماتریس بیانی و نرمال کردن داده‌ها، درصدهای مختلفی از گمشدگی، بر مجموعه داده‌های مورد مطالعه اعمال شد. سپس از ۱۱ الگوریتم بر پایه یادگیری ماشین به منظور بررسی تخمین داده‌های گمشده استفاده و میزان دقت هر یک از الگوریتمها، بر اساس نتایج حاصل مورد مقایسه قرار گرفت. بر اساس نتایج، دقت الگوریتمهای مختلف به مجموعه داده به کار رفته، درصد گمشدگی و توزیع گمشدگی داده‌ها وابسته است. همچنین تعداد نمونه‌های آزمایشی موجود در مجموعه داده‌ها نیز می تواند بر دقت الگوریتمهای تخمین داده‌های گمشده مؤثر باشد. نتایج بیانگر کاهش دقت تمامی الگوریتمها با افزایش درصد داده‌های گمشده بود، اما الگوریتمهای Least Square Adaptive و Local least square دقت بیشتری در مقابل افزایش درصد گمشدگی داده‌ها نسبت به سایر الگوریتمها نشان دادند.

واژه های کلیدی: الگوریتمهای بر پایه یادگیری ماشین، تخمین داده‌های گمشده، ریزآرایه

* نویسنده مسئول، تلفن: ۰۷۴۱-۲۲۲۴۸۴۰، پست الکترونیکی: mghaderi@yu.ac.ir

مقدمه

یکی از مشکلات جدی در مطالعات بیان ژن، به کارگیری داده‌های زیستی در حجمهای بالاست. نیاز به تولید، تحلیل و ادغام مجموعه داده‌های بزرگ و پیچیده زیستی منجر به پیشرفت روشهای پر توان (High Throughput) به منظور بررسی کل ژنوم مانند فناوری ریزآرایه شد. امروزه اهمیت استفاده از روشهای محاسباتی در دستیابی به نتایج دقیق‌تر بر کسی پوشیده نیست. این رویکردها عمدتاً به منظور تسریع تبدیل داده‌های مولکولی به اطلاعات معنی‌دار زیستی مورد استفاده قرار می‌گیرند (۲ و ۲۱). تجزیه و تحلیل داده‌های حاصل از فناوریهای پر توان، مانند ریزآرایه، به همراه تحلیل داده‌های آماری و بیوانفورماتیکی، برای کشف فرآیندهای پیچیده زیستی ضروری است (۹ و ۱۱). به طوری که این فناوری با استفاده از آشکارسازی نیمرخ بیان ژنها و طبقه‌بندی نمونه‌ها بر پایه الگوهای بیانی،

توانایی پاسخگویی به بسیاری از سئوالات ژنتیکی را دارد (۱۵).

داده‌های گمشده در اطلاعات ریزآرایه‌ها معمولاً طی آماده کردن اطلاعات به دلیل نقص در مراحل مختلف مانند دقت و تفکیک پذیری نامناسب، از بین رفتن تصویر، خراش یا گرد و خاک، وجود حباب بر روی اسلایدها و یا به صورت سیستماتیک در اثر روشهای ایجاد اسلایدها به وجود می‌آیند. متأسفانه به دلایل آزمایشگاهی و اقتصادی انجام دوباره آزمایشها مقرون به صرفه نیست. در زمان وجود داده‌های گمشده به طور معمول ساده‌ترین روش حذف بردار نیمرخ بیانی دارای داده گمشده و یا جایگزین کردن صفر (الگوریتم ZERO) یا میانگین ردیفها به‌جای داده‌های گمشده است (۲۴). طی سالهای گذشته الگوریتمهای مختلفی برای تخمین داده‌های گمشده، توسعه داده شده است که در ادامه نگاهی کوتاه بر الگوریتمهای استفاده شده در این مطالعه خواهد داشت.

اولین گزارش کاربرد الگوریتمهایی بر پایه یادگیری ماشین در تخمین داده‌های گمشده مربوط به الگوریتمهای K Nearest Neighbor (KNN)، Singular Value Decomposition (SVD) و Row Average (RAVG) است (۲۴)، سپس در سال ۲۰۰۳ الگوریتم دیگری به نام Bayesian Principle Component Analysis (BPCA) بر اساس روشهای آماری بیزین معرفی شد (۱۸ و ۲۳). پارامتری به نام K در الگوریتمهای SVD، KNN و Local Least Square (LLS) نیز بر دقت نتایج مؤثر است که معادل تعداد ژنهای ویژه‌ای (Eigengene) است که بیشترین شباهت را به ژن دارای داده گمشده دارند. انتخاب مقادیر K در BPCA و همچنین SVD به تعداد ژنهای اصلی بستگی دارد (۱۰). در روش LLS علاوه بر استفاده از مراحل بهینه‌سازی از طریق الگوریتم Local Square (LS) (الگوریتمی با خطای کم برای تخمین داده‌های گمشده بیان ژن)، از شباهتهای ساختارهای محلی نیز استفاده می‌شود (۱۶). الگوریتم Expectation Maximization (EM) برای هر دو تخمین بر اساس ژن و براساس آرایه، مورد استفاده قرار می‌گیرد (۶ و ۱۶). تخمین داده‌های گمشده مخلوط (داده‌هایی شامل گمشدگی کاملاً تصادفی Missing completely at Random (MCAR) و گمشدگی تصادفی Missing at Random (MAR)) از طریق الگوریتم MissForest اولین بار توسط روبین و همکاران (۱۹۷۸) ممکن شد (۷). این الگوریتم نسبت به KNN بدون توجه به ترکیب نوع متغیرها، ابعاد داده‌ها، منبع داده‌ها و یا میزان گمشدگی عملکردی بهتری دارد (۲۲). الگوریتم Sequential K-Nearest Neighbor (SKNN) از داده‌های تخمین زده شده برای تخمین داده‌های بعدی استفاده می‌کند. در این روش، داده‌های گمشده به صورت پشت سر هم از ژنهای دارای حداقل داده گمشده تخمین زده شده و برای محاسبات بعدی مورد استفاده قرار می‌گیرند (۱۷). روش Support Vector Regression (SVR) برای محاسبه چندین داده گمشده در هر ردیف نیمرخ بیانی از طرح ورودیهای مستقل (Orthogonal input coding scheme) استفاده می‌کند (۲۴). روش Least Square Adaptive (LSA) نیز از قانون حداقل مربعات استفاده می‌کند که در مقایسه با سایر الگوریتمها، دقت مناسب تری دارد (۶).

از آنجا که بسیاری از الگوریتمهای خوشه‌بندی و تعدادی از روشهای تجزیه و تحلیل آماری به مجموعه داده‌های کامل نیاز دارند، محاسبه داده‌های گمشده برای کاربرد مؤثر اطلاعات ریزآرایه امری ضروری است (۱۷). بنابراین برای به حداقل رساندن اثر مجموعه داده‌های ناقص جهت افزایش دامنه قابل اطمینان و همچنین تجزیه و تحلیل داده‌ها، از الگوریتمهای تخمین داده‌های گمشده بر پایه یادگیری ماشین استفاده می‌شود. هدف از این مطالعه مقایسه دقت الگوریتمهای تخمین داده‌های گمشده با تعداد مختلف ژنها و با درصدهای مختلف گمشدگی داده‌ها در سه مجموعه داده مختلف است.

مواد و روشها

سه مجموعه داده از سایت Gene Expression Omnibus (GEO) با پسوند CEL. دانلود شد (جدول ۱).

جدول ۱- مجموعه داده‌های به کار رفته جهت انجام تحلیلهای مورد مطالعه

ردیف	تعداد ژن × تعداد نمونه‌ها	چیپ به کار رفته برای آزمایش ریزآرایه	شماره دستیابی
۱	۶ × ۲۶۵۶۳۶	Affymetrix Porcine Genome Array	GSE32438
۲	۶ × ۲۶۵۶۲۸	Affymetrix Bovine Genome Array	GSE39796
۳	۷ × ۲۶۵۶۲۸	Affymetrix Bovine Genome Array	PMID: 20952064

الگوریتمهای به کار برده شده در این مطالعه به دو دسته الگوریتمهای Local imputation، الگوریتمهای Global imputation و الگوریتمهای دیگر SVR، EM و Missforset تقسیم شدند. الگوریتمهای Local imputation گروهی از ژنها با بیشترین ارتباط (فاصله اقلیدسی (۲۴))، همبستگی پیرسون (۶) و یا تخمین کوواریانس (۲۰) را برای محاسبه داده گمشده ژن هدف انتخاب می‌کند. برای الگوریتمهای Local از روشهای KNN، SKNN، LSA، LLS، Row average و Zero imputation استفاده شد. همچنین برای الگوریتمهای Global از روشهای SVD، BPCA استفاده شد (جدول ۲). برای ایجاد ماتریس کامل داده، تمام ژنهای دارای گمشدگی حذف شدند.

به منظور افزایش سرعت اجرای الگوریتمهای مختلف و همچنین امکان بررسی جزئیات عملکرد الگوریتمها در هر یک از مجموعه داده‌ها، از تعداد ۲۰۰ و ۸۰۰ ژن برای نمونه‌های مختلف آزمایشی استفاده شد. سپس برای یکسان‌سازی مقیاس اندازه‌گیری، هر سه مجموعه داده با استفاده از نرم افزار متلب، نسخه ۲۰۱۱ (MATLAB, version 2011) نرمال شدند. همچنین از تابع تولید گمشدگی (Miss generator Function)، در نرم افزار متلب، برای ایجاد درصدهای متفاوتی از گمشدگی در هر یک از مجموعه داده‌های نرمال شده و در نهایت بررسی اثر درصد داده‌های گمشده بر دقت تخمین الگوریتمهای مختلف، استفاده شد.

میزان دقت و کارآیی هر یک از الگوریتمها با استفاده از پارامتر آماری تابع خطای مربع میانگین ریشه نرمال شده (Normalized Root Mean Square Error (NRMSE)) محاسبه شد (فرمول ۱).

فرمول ۱

$$NRMSE = \frac{\sqrt{Mean\left[(y_{guess}-y_{ans})^2\right]}}{std\left[y_{ans}\right]}$$

NRMSE، معیاری برای تعیین تفاوت بین ارزشهای محاسبه شده و ارزش واقعی است که در آن y_{guess} و y_{ans} به ترتیب مقدار تخمین زده شده و مقدار واقعی داده‌ها هستند (۱۹). مناسب‌ترین الگوریتم تخمین، الگوریتمی است که کمترین میانگین، NRMSE را داشته باشد. به عبارت دیگر مقدار NRMSE بین صفر تا یک متغیر است که هرچه این مقدار به صفر نزدیک‌تر باشد، دقت الگوریتم بیشتر است. سپس از ۱۱ الگوریتم ZERO، RAVG، SVD، EM_gene، Missforset، BPCA، SKNN، SVR، KNN، LLS و LSA در هر سه مجموعه داده و با سه تکرار استفاده شد. به منظور نمایش دقت هر یک از الگوریتمها، تمام مقادیر تخمین زده شده داده‌های گمشده و الگوریتمهای به کار رفته با ۲۰۰ و ۸۰۰ ژن در سطوح گمشدگی ۵، ۱۰، ۱۵، ۲۰، ۲۵، ۳۰، ۴۵ و ۶۰ درصد بر اساس NRMSE مقایسه شدند.

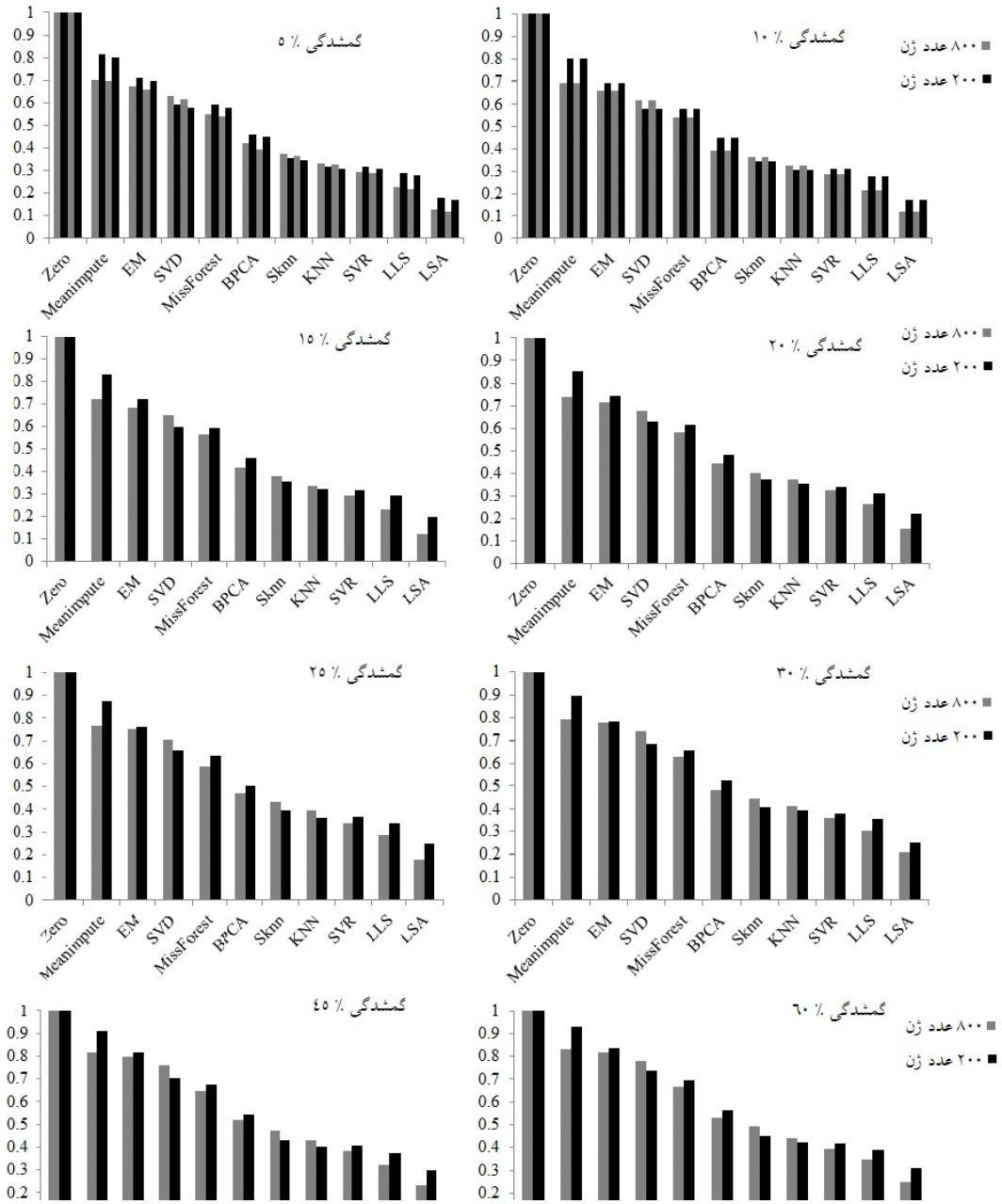
جدول ۲- خصوصیات مجموعه الگوریتمهای مورد استفاده برای تخمین داده گمشده

نام الگوریتم	روش کارکرد	مزایا	معایب	منبع
الگوریتمهای Local imputation				
Zero	قرار دادن صفر به جای داده‌های گم‌شده	محاسبه راحت	عدم استفاده از ساختار همبستگی داده‌ها، از بین رفتن قسمت مهمی از اطلاعات و دقت پایین ایجاد اختلال در توزیع ژنهای دارای داده گمشده، سخت شدن اندازه‌گیری‌ها، تخمین نادرست	Troyanskaya et al. (2001)
Row Average (RAVG)	استفاده از میانگین ردیفها به‌جای داده‌های گمشده	محاسبه راحت و سریع	انحراف معیار و درنتیجه کاهش قابلیت اطمینان به داده‌های حاصل و عدم استفاده از ساختار همبستگی داده‌ها	Troyanskaya et al. (2001)
Sequential K-Nearest Neighbor (SKNN)	تخمین داده‌های گمشده بر پایه خوشه‌بندی	حفظ ساختار اصلی داده‌ها، مناسب برای داده‌هایی با نسبت زیاد گمشدگی و تعداد زیاد آزمایشها، ارزش بالای نتایج حاصل از خوشه‌بندی و مقاوم در برابر تخمین مجموعه داده‌هایی با توزیع ناساوی گمشدگی	پیچیدگی محاسباتی زیاد	Kim et al. (2004)
K Nearest Neighbor (KNN)	استفاده از K ژن با بیشترین شباهت به ژنهای دارای داده گمشده بر اساس میانگین وزنی	عملکرد مناسب برای داده‌هایی با همبستگی بالا و پیچیدگی زیاد	مشکل درانتخاب مقدار پارامتر K	Troyanskaya et al. (2001)
LLS	نسخه بهبودیافته روش KNN با بکارگیری روش حداقل مربعات بجای وزنهای پیشنهاد شده	دقت مناسب در داده‌هایی با پیچیدگی زیاد	مشکل تعیین تعداد مناسب پارامتر K برای ژنهایی با بیشتر از یک داده گمشده	Kim et al. (2005)
Least Square Average (LSA)	تخمین داده‌های گمشده بر اساس رگرسیون به منظور نمایش ارتباط بین ژنها و آرایه‌ها	نمایش ارتباط بین خطاهای مجموع array، قابلیت انعطاف برای تنظیم تعداد ژنهای مؤثر در رگرسیون، در نظر گرفتن رگرسیون منفی و مثبت ژنهای مرتبط، سرعت و دقت بالا در تخمین داده‌های گمشده در	LS-gene و LS- دقت کم تخمین در داده‌هایی که باهم ارتباط غیر خطی دارند	Bo et al. (2004)

		دادههایی با پیچیدگی زیاد		
الگوریتم‌های Global imputation	Support Vector Regression (SVR)	تخمین دادههای گمشده برپایه قانون حداقل احتمال ساختاری	بیشترین استفاده از اطلاعات اصلی شبکه داده و سبب بهبود عملکرد مناسب تخمین دادههای گمشده توجه به تمامی متغیرهای مؤثر	مشکل درانتخاب مقدار پارامتر K، مشکل در تخمین تابع مناسب کرنل و پیچیدگی محاسباتی
	Expectation Maximization (EM)	براساس مدل رگرسیون چندگانه	این الگوریتم تا زمان ثابت شدن تخمین مقادیر دادههای گمشده و ماتریس کوواریانس تکرار می‌شود. درنتیجه همگرایی تدریجی کوواریانس و تخمین دادههای گمشده بوجود می‌آید	سرعت نسبتا پایین
	Missforset	تخمین دادههای گمشده بر اساس روش غیرپارامتری با به کارگیری همزمان انواع مختلف متغیرها	عملکرد مناسب در دادههایی با روابط پیچیده و غیرخطی، تخمین دادههایی با ابعاد بالا، استفاده همزمان از متغیرهای پیوسته، گسسته و مخلوط و عدم نیاز برای مقداردهی به پارامترها	دقت تخمین کم در مجموعه دادههایی با روابط خطی
	Bayesian Principle Component Analysis (BPCA)	تخمین دادههای گمشده براساس تئوری احتمالات بیزی	سرعت بالا با وجود پیچیدگی محاسباتی	دقت تخمین نسبتا کمی و مناسب برای دادههایی با پیچیدگی کم و دقت پایین‌تر در دادههایی با ساختار همبستگی
	Singular Value Decomposition (SVD)	استفاده دسته‌ای از الگوهای بیانی مشترک ماتریس بیان ژن	مناسب برای دادههایی که ساختار همبستگی ندارند	نمایش تمامی ژنها در ماتریس بیان، عدم کاربرد برای ژنهایی با الگوی بیان متفاوت و مناسب برای دادههایی با پیچیدگی کم دقت پایین‌تر در دادههایی با ساختار همبستگی
				Wang et al. (2006)
				Kim et al. (2005)
				Daniel J. Stekhoven (2011)
				Oba et al. (2003)
				Troyanskaya et al. (2001)

نتایج و بحث

امروزه دانشمندان برای بررسی بیان ژن از روشهای مختلفی (مانند فناوری ریزآرایه، RT-PCR نیمه کمی و غیره (۱ و ۴)، در تحقیقات مختلف و در موجودات متفاوت (مانند سرطان خون (۴) و آدنوکارسینومای معده انسان (۳)، گیاه ریحان (*Ocimum basilicum* L.) (۱) و غیره) استفاده می-کنند. در این مطالعه، سه مجموعه داده ریزآرایه و ۱۱ الگوریتم تخمین داده گمشده به کار گرفته شدند. دقت تمامی الگوریتمها با افزایش درصد گمشدگی، کاهش یافت (شکل ۱).

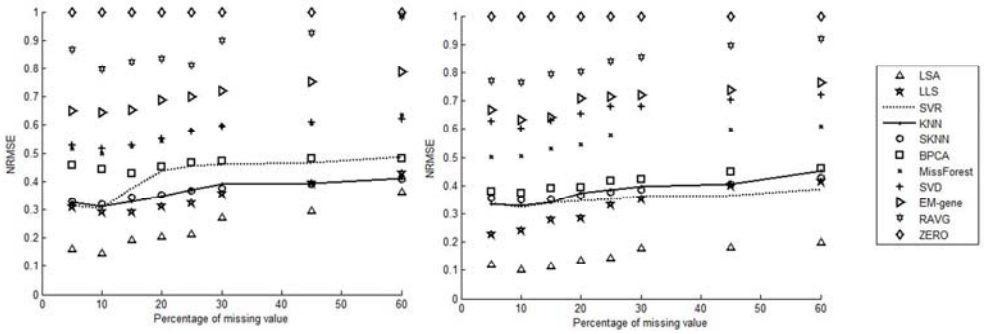


شکل ۱- مقایسه میانگین دقت روشهای تخمین داده‌های گمشده با ۵ درصد داده گمشده در سه مجموعه داده مورد مطالعه

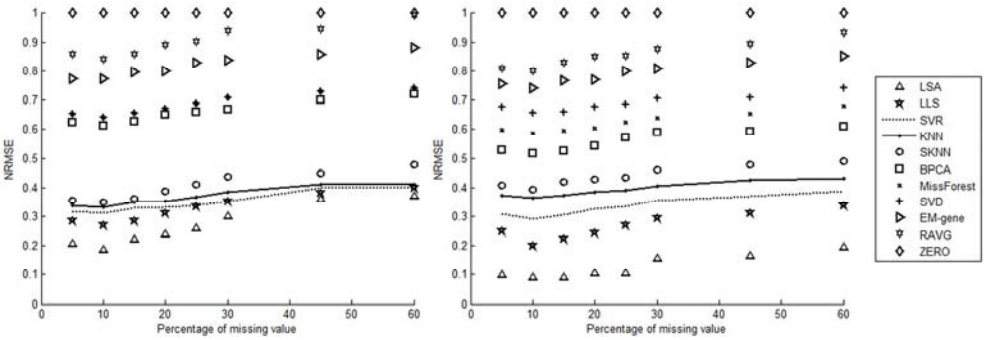
همچنین عملکرد و دقت هر کدام از الگوریتمها به تعداد ژنها، نمونه‌ها و خصوصیات مجموعه داده مورد مطالعه وابسته بود. بر این اساس دقت و عملکرد الگوریتمهای ZERO و RAVG، به درصد گمشدگی و توزیع گمشدگی داده‌های گمشده بستگی دارد و کمترین میزان دقت تخمین داده‌های گمشده از این دو الگوریتم به دست آمد.

روشهای جایگزینی داده‌های گمشده با صفر و یا میانگین ردیفها (RAVG) از ساختار همبستگی داده‌ها استفاده نمی-کنند. به طور پیش‌فرض دقت روش تخمین جایگزینی صفر به جای داده گمشده، همیشه یک (یک برابر است با بیشترین مقدار NRMSE، به عبارت دیگر کمترین میزان دقت) است (۲۴). به طوری که، پس از روش ZERO، دقت و عملکرد الگوریتم RAVG در تخمین داده‌های گمشده موجود در تمامی مجموعه‌داده‌ها، برای ۲۰۰ ژن (شکلهای ۲ تا ۴، سمت چپ) و ۸۰۰ ژن (شکلهای ۲ تا ۴، سمت راست) کمترین میزان نسبت به سایر الگوریتمها بود.

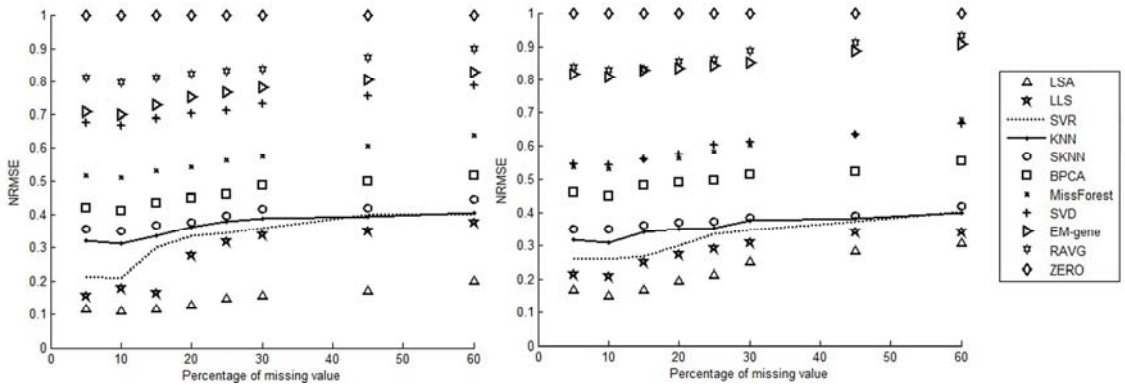
الگوریتم MissForest در مجموعه داده‌های یک، دو و سه، دقت مناسبی برای تخمین داده‌های گمشده نداشت. با این وجود دقت تخمین آن از الگوریتمهای SVD، RAVG و EM_gene بهتر بود. همچنین با کاهش تعداد ژنهای مورد بررسی (از ۸۰۰ ژن به ۲۰۰ ژن) دقت تخمین این الگوریتم نیز کاهش یافت. البته دقت الگوریتم‌های SDV و KNN بیشتر از روشهای جایگزینی صفر به جای داده گمشده و یا قرار دادن میانگین درایه‌های مشابه به ژن دارای داده گمشده (RAVG)، است. در این روش بر اساس فرض شباهت بیان یک ژن در یک آزمایش با میانگین بیان ژنها در زمانهای مختلف انجام آزمایش استوار است، به همین دلیل دقت پایین این روش امری دور از انتظار نیست (۲۴).



شکل ۲- مقایسه الگوریتمهای مختلف با گمشدگیهای متفاوت در مجموعه داده ۱ (سمت چپ: ۲۰۰ ژن و سمت راست: ۸۰۰ ژن)



شکل ۳- مقایسه الگوریتمهای مختلف با گمشدگیهای متفاوت در مجموعه داده ۲ (سمت چپ: ۲۰۰ ژن و سمت راست: ۸۰۰ ژن)



شکل ۴- مقایسه الگوریتمهای مختلف با گمشدگیهای متفاوت در مجموعه داده ۳ (سمت چپ: ۲۰۰ ژن و سمت راست: ۸۰۰ ژن)

در الگوریتم KNN با افزایش ژنها میزان NRMSE نیز بیشتر می‌شود. این الگوریتم نسبت به افزایش درصد داده-های گمشده مقاوم و کاهش دقت آن در مقابل افزایش درصدهای گمشدگی کمتر از سایر الگوریتمها بود، به طوری که با ۲۰ درصد داده گمشده، حداکثر ۱۰ درصد کاهش دقت را نشان داد (۲۴). همچنین الگوریتم KNN برای محاسبه داده‌های گمشده برای ژنهایی که در خوشه-های کوچک بیان می شوند، دقیق تر است (۲۴). بنابراین دقت این الگوریتم برای ۲۰۰ ژن از ۸۰۰ ژن در هر سه مجموعه داده مورد مطالعه بیشتر بود و میزان NRMSE کوچکتری به دست آمد (شکلهای ۲ تا ۴). در این الگوریتم با کاهش تعداد ژنها، دقت بیشتری برای تخمین داده‌های گمشده به دست آمد. الگوریتم KNN برای پروفایلهای بیانی با تعداد کم ژنها، نتایج دقیق‌تری دارد، به علاوه رابطه کاهش دقت آن با افزایش درصد گمشدگی، معنی دار نیست (۸ و ۱۱).

نتایج حاصل از تخمین SVD نشان دادند که این الگوریتم دقت کمتری نسبت به سایر روشها دارد و با کاهش تعداد ژنها از ۸۰۰ به ۲۰۰ ژن، مقدار NRMSE به دست آمده از این روش نیز در هر سه مجموعه داده کاهش یافت (شکلهای ۲ تا ۴). بر اساس خصوصیات داده‌ها، دقت تخمین الگوریتمهای MissForset و SVD تقریباً در دامنه ۰/۵ تا ۰/۶۳ بود. الگوریتم EM_gene نیز دقت پایینی (۰/۶۴) را نشان داد.

قدرت تخمین الگوریتم BPCA به تعداد نمونه‌ها یا ژنهای مورد بررسی بستگی دارد. در این مطالعه در هر سه مجموعه داده به دلیل تعداد کم نمونه‌های مورد بررسی (شش و هفت نمونه) دقت الگوریتم BPCA کمتر از KNN و حتی روش SKNN بود. Oba و همکارن (۲۰۰۳) نشان دادند که دقت الگوریتم BPCA با افزایش تعداد ژنها (حداقل ۴۰ نمونه) رابطه مستقیم دارد (۱۸). همچنین دقت الگوریتم BPCA نیز با افزایش درصد گمشدگی رابطه معکوس داشت که این امر توسط Celton و Malpertuy (۲۰۱۰) نیز گزارش شده است (۸).

نتایج حاصل از مقایسه‌ها در داده‌هایی با ۵ درصد گمشدگی نشان دادند که میزان NRMSE حاصل از الگوریتم LSA، ۱۵ تا ۲۰ درصد کوچکتر از نتایج حاصل از الگوریتم KNN است. با کاهش تعداد ژنها، دقت به دست آمده از الگوریتم LSA نیز به دلیل استفاده داده‌ها و روابط همبستگی کمتر، کاهش یافت. ارزشهای NRMSE در الگوریتم SKNN بیشتر از BPCA است. همچنین دقت الگوریتم SKNN به خصوص با افزایش درصد گمشدگی، بیشتر از الگوریتم EM_gene است. بر اساس نتایج هر سه مجموعه داده افزایش تعداد ژنها سبب کاهش دقت الگوریتم SKNN شد.

تخمین LLS میزان NRMSE کمتری نسبت به KNN و BPCA با ۵ تا ۶۰ درصد داده گمشده نشان داد. در این

الگوریتم با کاهش ژنهای مورد مطالعه (از ۸۰۰ ژن به ۲۰۰ ژن)، خطای محاسباتی نیز بیشتر شد. دقت الگوریتم LLSدر تخمین داده‌های گمشده بیشتر از الگوریتمهای KNN، BPCA و SVD گزارش شده است (۱۲). همچنین الگوریتم SVR عملکرد ثابتی در درصدهای مختلف گمشدگی در هر سه مجموعه داده داشت. با افزایش مقدار گمشدگی، افزایش NRMSE حاصل از الگوریتم SVR اندکی بیشتر از مقدار این مقیاس برای الگوریتمهای LLS و BPCA و بسیار کمتر از الگوریتم KNN بود. هنگامی که درصد داده‌های گمشده بسیار زیاد باشد، محاسبه SVR در مقایسه با BPCA و LLS به دلیل رویکردهای شبکه‌ای برای مجموعه پارامترها، عملکرد ضعیف‌تری را نشان خواهد داد (۲۵). الگوریتم SVR در مجموعه داده‌هایی با همبستگی بالای ژنی، عملکرد بهتری از الگوریتمهای KNN و BPCA داشت. با کاهش تعداد ژنهای مورد استفاده برای تخمین در هر سه مجموعه داده، میزان NRMSE در الگوریتم SVR بیشتر شد و در نتیجه دقت کاهش یافت.

به طور کلی داده‌های گمشده مشکلاتی را برای تجزیه و تحلیل مجموعه داده‌های حاصل از آزمایشهای ریزآرایه به وجود می‌آورند. بنابراین نیاز به برطرف کردن این مشکلات امری ضروری است (۲۴). اولین راه حل ممکن، کم کردن حجم مجموعه داده‌ها از طریق حذف ژنهای دارای داده گمشده است. این روش برطرف کردن داده‌های گمشده هنوز در عمل توسط بسیاری از محققین مورد استفاده قرار می‌گیرد (۱۲ و ۱۴). علی رقم وجود روشهای مختلفی برای تخمین داده‌های گمشده، روشهای جدید توسعه یافته‌تر هستند. البته روشهایی که از آنها به طور گسترده‌ای در تحلیل داده‌های ریزآرایه استفاده می‌شود، اغلب سبب اریب شدن و به غلط انداختن محاسبه داده‌های گمشده شوند. بر این اساس توافق عمومی برای چگونگی انتخاب روشهای مختلف وجود ندارد، چون به نظر می‌رسد که عملکرد هر یک از آنها به مقدار زیادی به مجموعه داده مورد بررسی، درصد داده‌های گمشده، توزیع گمشدگی داده‌ها، وجود عوامل اضافی مانند نویز و غیره وابسته است (۲۶). برای دستیابی به بهترین روش تحلیل داده گمشده و به دست آوردن ماتریس کامل بیان ژن، باید داده‌های گمشده و مناسب‌ترین الگوریتمهای تخمین داده‌های گمشده در مجموعه داده‌های مختلف با شرایط متفاوت، مورد بررسی و شناسایی قرار گیرند (۱۳). به طور کلی نتایج حاصل از مقایسه الگوریتمهای مختلف در منابع متفاوت، حاکی از این حقیقت است که هیچ الگوریتم تخمین مطلوبی برای تمامی انواع داده‌ای وجود ندارد. با این وجود و در این مطالعه در میان الگوریتمهای مورد بررسی در هر سه مجموعه داده، الگوریتم LSA بیشترین دقت و قدرت تخمین و همچنین کمترین NRMSE را در تمامی مجموعه داده‌های مورد مطالعه به خود اختصاص داد (شکلهای ۲ تا ۴). این امر با ساختار همبستگی محلی (Local Correlation) داده‌ها و شباهت ژنهای مورد مطالعه و همچنین عملکرد روش LSA مرتبط است. عملکرد روش LSA براساس قانون حداقل مربعات و استفاده همزمان از ارتباط بین ژنها و آرایه‌ها تعریف می‌شود. در این روش قانون حداقل مربعات براساس حداقل کردن مجموع خطاهای مربع مدل مورد بررسی است. علت استفاده از ارتباط ژنها به عنوان اساس تخمین، پدیده تنظیم همزمان ژنها طی فرآیندهای عملکردی سلول است، مانند نقش تنظیم کنندگی ژن *ein2* در گیاه اطلسی بر مسیر انتقال پیام اتیلن (۵). از طرف دیگر استفاده از پروفایلهای بیانی گرفته شده از آرایه‌های مختلف نیز از اهمیت به سزایی برخوردار است چرا که هیبریداسیون‌های آرایه‌ای نمونه‌های بیولوژیکی حاصل از بافتهای یکسان معمولاً با یکدیگر در ارتباط هستند، بر این اساس انتظار می‌رود اندازه ستونهای مختلف آنها در ماتریس بیانی ژنها نیز یکسان باشد. اگرچه عکس این مطلب نیز صادق است (نمونه‌های زیست‌شناسی بسیار متفاوت، پایه و اساس ضعیفی برای تخمین ماتریس بیانی هستند). همان طور که پیش‌تر نیز اشاره شد الگوریتم

LSA مانند KNN از میانگین وزنی ژنهایی با همبستگی بیشتر به ژنهای دارای داده گمشده استفاده می‌کند و سپس با استفاده از معادله حداقل مربع، داده گمشده را محاسبه می‌کند. البته میانگین وزنی در الگوریتم LSA بر اساس روش وزنی سازگار (Adaptive Weighting Scheme) و با در نظرگیری ساختار همبستگی داده‌ها تعیین می‌شود.

الگوریتمهای LLS و SVR نیز در درجه بعد، تخمین قوی-تری را نشان دادند و کمترین قدرت و دقت در الگوریتمهای RAVG و ZERO به دست آمد.

منابع

۱. تحصیلی، ژ. شریفی، م. بهمنش، م. ضیایی، م. ۱۳۸۹. بیان ژن آنزیم اوژنول-O – متیل ترانسفراز و ارتباط آن با اجزاء اسانس در مراحل مختلف رشد ریحان (*Ocimum basilicum* L.). مجله زیست‌شناسی ایران. (۱) ۲۳. ص: ۲۵-۱۸.

۲. خلیلی، س. جهانگیری، الف. امانی، ج. سلمانیان، ع. ه. ۱۳۹۳. کاربرد بیوانفورماتیک در مطالعات ایمنی شناسی. مجله زیست-شناسی ایران. (۲) ۲۷. ص: ۲۱۰- ۱۹۲.

۳. رنجی، ن. پادگانه، الف. صادقی‌زاده، د. صادقی‌زاده، م. ۱۳۹۳. بررسی بیان ژنهای *hTERT* و *Survivin* در رده سلولی آدنوکارسینومای معده انسان (AGS) تحت تیمار با نانوکورکومین. مجله زیست‌شناسی ایران. (۲) ۲۷. ص: ۲۴۱-۲۳۳.

۴. قلندری، م. بهمنش، م. اکبری، م. ت. ۱۳۹۱. بررسی بیان ژن SUZ12 به عنوان مارکری از تغییرات اپی ژنتیکی در بیماران مبتلا به سرطان خون (CML). مجله زیست‌شناسی ایران. (۲) ۲۵. ص: ۲۲۹- ۲۲۰.

۵. میرشمسی کاخکی، الف. بهرامی، الف. ر. شهریاری احمدی، ف. گری، ج. ۱۳۹۲. بررسی بیان ژن *ein2* در گیاه اطلسی (Petunia×hybrida) و مطالعه نقش تنظیم کنندگی آن در مسیر انتقال پیام اتیلن. مجله زیست‌شناسی ایران. (۴) ۲۶. ص: ۵۸۷-۵۷۲.

12. Horton, N.J. and Kleinman, K.P. 2007. Much ado about nothing, The American Statistician, 61.
13. Hourani, M.A. and El Emary, I.M. 2009. Microarray missing values imputation methods: Critical analysis review, Computer Science and Information Systems/ComSIS, 6: 165-190.
14. Kaiser, J. 2012. Algorithm for Missing Values Imputation in Categorical Data with Use of Association Rules, ACEEE International Journal on Recent Trends in Engineering & Technology.
15. Kerr, M.K., Martin, M. and Churchill, G.A. 2000. Analysis of variance for gene expression microarray data, Journal of computational biology, 7: 819-837.
16. Kim, H., Golub, G.H. and Park, H. 2005. Missing value estimation for DNA microarray gene expression data: local least squares imputation, Bioinformatics, 21: 187-198.
17. Kim, K.-Y., Kim, B-J. and Yi, G.-S. 2004. Reuse of imputed data in microarray analysis increases imputation efficiency, BMC bioinformatics, 5: 160.

6. Bø, T.H., Dysvik, B. and Jonassen, I. 2004. LSimpute: accurate estimation of missing values in microarray data with least squares methods, Nucleic acids research, 32: e34-e34.
7. Breiman, L. 2001. Random forests, Machine learning, 45: 5-32.
8. Celton, M., Malpertuy, A., Lelandais, G. and De Brevern, A.G. 2010. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments, BMC genomics, 11: 15.
9. Fellenberg, K., Busold, C.H., Witt, O., Bauer, A., Beckmann, B., Hauser, N.C., Frohme, M., Winter, S., Dippon, J. and Hoheisel, J.D. 2006. Systematic interpretation of microarray data using experiment annotations, BMC genomics, 7: 319.
10. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R. and Caligiuri, M.A. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, 286: 531-537.
11. Hoheisel, J.D. 2006. Microarray technology: beyond transcript profiling and genotype analysis, Nature reviews genetics, 7: 200-210.

18. Oba, S., Sato, M.-a., Takemasa, I., Monden, M., Matsubara, K.-i. and Ishii, S. 2003. A Bayesian missing value estimation method for gene expression profile data, Bioinformatics, 19: 2088-2096.

19. Ouyang, M., Welsh, W.J. and Georgopoulos, P., 2004. Gaussian mixture clustering and imputation of microarray data, Bioinformatics, 20: 917-923.

20. Sehgal, M.S.B., Gondal, I. and Dooley, L.S. 2005. Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data, Bioinformatics, 21: 2417-2423

21. Stekel, D. 2003. Microarray bioinformatics. Cambridge University Press.

22. Stekhoven, D.J. and Bühlmann, P. 2012. MissForest—non-parametric missing value imputation for mixed-type data, Bioinformatics, 28: 112-118.

23. Takemasa, I., Higuchi, H., Yamamoto, H., Sekimoto, M., Tomita, N., Nakamori, S., Matoba, R., Monden, M. and Matsubara, K. 2001. Construction of preferential cDNA microarray specialized for human colorectal carcinoma: Molecular sketch of colorectal cancer, Biochemical and biophysical research communications, 285: 1244-1249.

24. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. 2001. Missing value estimation methods for DNA microarrays, Bioinformatics, 17: 520-525.

25. Wang, X., Li, A., Jiang, Z. and Feng, H. 2006. Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme, BMC bioinformatics: 7, 32.

26. Yeung, K.Y., Medvedovic, M. and Bumgarner, R.E. 2003. Clustering gene-expression data with repeated measurements, Genome Biol, 4: R34.

Comparison of accuracy of machine learning algorithms on missing values estimation of DNA microarray datasets

Moshiry M.[1], Ghaderi-Zefrehei M.[2] and Ghane Golmohammadi F.[3]

[1] Animal Science Dept., Agriculture Faculty, Ferdowsi University of Mashhad, Mashhad, I.R. of Iran

[2] Animal Science Dept., Agriculture Faculty, Yasouj University, Yasouj , I.R. of Iran

[3] Systems Biology Dept., Agricultural Biotechnology Research Institute of Iran, Karaj, I.R. of Iran

Abstract

Presence of missing values in microarray data decreases accuracy of drawing regulatory gene networks and may cause mistake in clustering and specialized classification of genes and further analysis. Therefore, missing values estimation is one of the most important steps in preprocessing of microarray data. Function of estimation algorithms is varied in different datasets and different missing percentage. Select a proper algorithm, in order to achieve the most accurate results in calculation of missing values, is a critical point. In this study, three microarray datasets were used. Dimensions of expression matrix was determined, and data normalization was carried out, then, different missing percentages were applied on under studied datasets. 11 machine learning algorithms were used to estimate the missing values, and their accuracy were compaied based on the output. Based on the archeived results, accuracy of each algorithms depends on used datasets, missing percentage, and missing distribution. Also, the number of experimental samples in datasets can affect the accuracy of missing values estimation algorithms. The results revealed a descending trend in accuracy over increasing the percentage of missing data. However, Least Square Adaptive and Local Least Square algorithms showed more accuracy through increasing of missing percentage rather than others.

Key words: Machine Learning Algorithms, Missing value estimation, Microarray