ارزیابی صحت گروه‌بندی روش‌های مختلف تجزیه خوشه‌ای

مهدی رمضانی[۱*]، مهدی رحیمی[۱]، حبیب اله سمیع زاده لاهیجی[۲] و فهیمه رحیمی[۳]

چکیده

در این مطالعه، میزان صحت گروه‌بندی ۱۸ ژنوتیپ گیاه ذرت توسط روش‌های مختلف تجزیه خوشه‌ای، بر پایه روش‌های مختلف استاندارد کردن داده‌ها و شاخص‌های متفاوت فاصله‌ای، با تجزیه تابع تشخیص مورد ارزیابی قرار گرفت. بدین منظور ۷ لاین اینبرد و ۷ هیبرید حاصل از تلاقی آن‌ها به همراه ۴ توده حاصل از گرده‌افشانی آزاد هیبریدها در قالب یک طرح بلوک‌های نامتعادل گروهی در سه تکرار کشت گردید و ۲۲ صفت بر روی این ژنوتیپ‌ها اندازه‌گیری شد. تجزیه تابع تشخیص نشان داد که شاخص فاصله اقلیدسی بهتر از سایر شاخص‌های فاصله‌ای بود و گروه‌بندی مطلوبی بر اساس آن حاصل گردید. هم‌چنین هر سه روش استاندارد کردن داده‌ها، گروه‌بندی یکسانی داشتند و نتایج آن‌ها بهتر از استاندارد نکردن داده‌ها بودند. ارزیابی دندروگرام‌های روش‌های مختلف تجزیه خوشه‌ای نشان داد که روش‌های متوسط فاصله بین دسته‌ها، دورترین همسایه‌ها و حداقل واریانس "وارد" بهتر از سایر روش‌ها بودند. تجزیه تابع تشخیص خطی فیشر نیز نشان داد که روش‌های متوسط فاصله بین دسته‌ها، دورترین همسایه‌ها و حداقل واریانس "وارد" با انجام صحت گروه‌بندی در حدود ۱۰۰٪، بهتر از سایر روش‌های دیگر توانسته‌اند از اطلاعات اولیه داده‌ها استفاده و ژنوتیپ‌ها را گروه‌بندی نمایند.

واژهای کلیدی: تجزیه تابع تشخیص، تجزیه خوشه‌ای، ذرت

۱ و ۲. به ترتیب دانشجویان کارشناسی ارشد و استادیار گروه زراعت و اصلاح نباتات، دانشکده علوم کشاورزی دانشگاه گیلان
۳. دانشجوی کارشناسی ارشد برنامه‌ریزی و مدیریت محیط زیست، دانشکده محیط زیست دانشگاه تهران
*: نویسنده مسوول

مقدمه

ذرت یکی از محصولات مهم در ایران است که سطح کشت و عملکرد آن در دهه اخیر به طور چشم-گیری افزایش یافته است. چنین افزایشی در سطح کشت و همراه با آن، افزایش عملکرد در واحد سطح نیازمند یک برنامه اصلاحی موثر است تا بتوان از پتانسیل هتروزیس بین ژرم‌پلاسم‌های تولیدی استفاده نمود. برای آن‌که به‌نژادگر بتواند حداکثر بهره‌برداری را از پدیده هتروزیس به عمل آورد، ابتدا لازم است میزان تنوع موجود در بین ژنوتیپ‌های مورد مطالعه را ارزیابی نماید وسپس با دورگ‌گیری بین ژنوتیپ‌هایی که از نظر تنوع، تفاوت عمده‌ای با یکدیگر دارند به هیبریدهای پر محصول و با صفات مطلوب دست یابد. تجزیه خوشه‌ای یکی از روش‌های آماری چند متغیره است که برای تعیین تنوع بین جوامع مختلف گیاهی، جانوری و نیز دسته‌بندی آن‌ها به گروه‌های مختلف بر اساس فاصله ژنتیکی و یا تشابه ژنتیکی به‌کار گرفته می‌شود (رومسبرگ، ۱۹۹۰). این روش حداقل در دو مورد می-تواند به به‌نژادگر کمک نماید: یکی پیدا کردن گروه‌های واقعی افراد بر اساس تشابه ژنتیکی بین آن‌ها و دیگر کاهش داده‌ها و انتخاب افراد محدودی از هر گروه یا دسته (جابسن، ۱۹۹۲). گروه‌بندی ژرم‌پلاسم‌های ذرت بر اساس صفات مورفولوژیک به کرات توسط پژوهش-گران مختلف دنیا انجام گرفته است (بریتینگ و همکاران، ۱۹۹۰ و کروسا و همکاران، ۱۹۹۵). کارآیی شاخص تشخیص را هم می‌توان به‌وسیله تخمین احتمالات کلاس‌بندی نادرست از مشاهدات جدید در ارزیابی داده‌ها ارزیابی کرد (فرناندز، ۲۰۰۶). مندز و همکاران (۲۰۰۲) با استفاده از تجزیه تشخیص بر روی ۹ گروهی که از تجزیه خوشه‌ای با روش متوسط فاصله بین و درون کلاسترها با شاخص فاصله‌ای پیرسون به دست آمده بود، نشان دادند که از ۹ گروه اولیه، ۵ گروه در حدود ۶۰٪ درست گروه‌بندی شده بودند و فقط دو گروه ۱۰۰٪ صحیح گروه‌بندی شده بودند. موردا و همکاران (۲۰۰۳) نیز با آنالیز تجزیه خوشه‌ای به روش حداقل واریانس وارد و شاخص فاصله اقلیدسی، ۸۵ نمونه چای را در دو گروه آسیایی و آفریقایی گروه‌بندی نمودند و سپس با آنالیز تابع تشخیص نشان دادند که ۹۴/۴٪ از این گروه‌بندی، صحیح انجام شده بود. در پژوهشی هم که جینز و همکاران (۲۰۰۳) بر روی داده-های مزرعه‌ای ذرت انجام دادند، ژنوتیپ‌های مورد مطالعه را در ۵ گروه دسته‌بندی کردند و با تابع تشخیص به روش کنارگذاری نشان دادند که ۸۰٪ از گروه‌بندی‌ها صحیح انجام گرفته بود.

اما مشکل عمده و اساسی استفاده از روش تجزیه خوشه‌ای در گروه‌بندی افراد یا ژنوتیپ‌ها این است که روش‌های بسیار متفاوتی برای انجام این نوع تجزیه توسط پژوهشگران مختلف پیشنهاد شده است که عمدتاً نتایج کاملاً متفاوتی نیز ارائه می‌دهند. به این ترتیب تشخیص صحت و سقم نتایج حاصل از روش‌های مختلف و گروه‌بندی‌های حاصل به وسیله پژوهش‌گر، بسیار مشکل و گمراه کننده خواهد بود. یکی از روش‌هایی که از آن برای تشخیص صحت گروه‌بندی به‌دست آمده از تجزیه‌ی خوشه‌ای می‌توان استفاده نمود، تجزیه تابع تشخیص است که برای این منظور می‌توان از روش‌های کنارگذاری و اعتباری استفاده نمود (جابسن، ۱۹۹۲ و جیانگ، ۲۰۰۱).

در این پژوهش از داده‌های اندازه‌گیری شده‌ی ارزش‌های فنوتیپی ۲۲ صفت مهم در ۷ لاین، ۷ هیبرید و ۴ توده حاصل ازهیبریدهای ذرت، برای انجام گروه-بندی اولیه ژنوتیپ‌ها با روش‌های مختلف تجزیه خوشه-ای، شاخص‌های متفاوت فاصله و روش‌های مختلف استاندارد کردن داده‌ها استفاده شد. سپس گروه‌های حاصل با تجزیه تابع تشخیص مورد ارزیابی قرار گرفت تا میزان صحت گروه‌بندی‌های حاصل از شاخص‌ها و روش‌های مختلف تجزیه خوشه‌ای مورد بحث و بررسی قرار گیرند.

مواد و روش‌ها

برای انجام این آزمایش تعداد ۷ لاین والدی، ۷ هیبرید حاصل از تلاقی والدین به همراه توده‌های حاصل از گرده‌افشانی آزاد هیبریدهای SC 301، SC 604، SC 647 و SC 704 (جدول ۱) در قالب طرح بلوک‌های نامتعادل گروهی در ۳ تکرار در بهار ۱۳۸۴ در مرکز

تحقیقات کشاورزی همدان کشت گردید. زمین مورد آزمایش در پاییز سال قبل شخم خورده و در بهار پس از انجام عملیات خاک‌ورزی ثانویه آماده کشت گردید. با توجه به آزمایش خاک تنها از کود اوره به میزان ۳۰۰ کیلوگرم در هکتار به صورت سرک در ۳ مرحله استفاده گردید. هر تیمار در ۴ خط کشت با فواصل بین ردیف ۷۵ سانتی‌متر و داخل ردیف ۲۵ سانتی‌متر کشت گردیدند.

صفات اندازه‌گیری شده عبارت از: تعداد روزها تا ظهور ۵۰ ٪ کاکل، تعداد روزها تا ظهور ۵۰٪ تاسل، تعداد روزها تا خشک شدن کاکل، تعداد روز تا ظهور گل تاجی، تعداد روز تا خاتمه گرده افشانی، تعداد روزها تا رسیدگی فیزیولوژیک، ارتفاع بوته، ارتفاع تاسل، تعداد گره، متوسط طول میان‌گره، طول، عرض، مساحت برگ بلال اصلی، تعداد کل برگ در بوته، تعداد برگ بالای بلال اصلی، ارتفاع محل بلال، طول بلال، تعداد ردیف دانه، تعداد دانه در ردیف، عمق دانه، عملکرد دانه در بوته و وزن هزار دانه بودند. برای انجام تجزیه خوشه‌ای نیز از میانگین داده‌های هر ژنوتیپ استفاده گردید. ابتدا از داده‌های اصلی و استاندارد نشده استفاده گردید. سپس داده‌ها به سه روش آماری (روابط ۱، ۲ و ۳) استاندارد شدند و به این ترتیب، چهار گروه از داده‌ها برای انجام تجزیه خوشه‌ای به‌کار رفتند.

رابطه (۱)
$$z = \frac{x_i - \bar{x}}{s_x}$$

رابطه (۲)
$$\alpha_i^* = \frac{x_i}{x_{Max} - x_{Min}}$$

رابطه (۳)
$$\alpha_i^* = \frac{x_i - x_{Min}}{x_{Max} - x_{Min}}$$

از شاخص‌های فاصله اقلیدسی، پیرسون، مینکوسکی[1]، چبی‌چف[2]، کوساین[3] و بلوک[4] برای تعیین فاصله بین ژنوتیپ‌ها استفاده شد (رومسبرگ، ۱۹۹۰). جهت ارزیابی روش‌های مختلف تجزیه خوشه‌ای نیز از هفت روش تجزیه مختلف شامل متوسط فاصله بین کلاسترها، دورترین همسایه‌ها، متوسط فاصله بین و درون کلاسترها، نزدیک‌ترین همسایه‌ها، مرکزی، میانه‌ای و حداقل واریانس وارد استفاده شد. به منظور تعیین تعداد خوشه‌ها نیز از روش بیش‌ترین گسیختگی با توجه به سادگی آن استفاده شد که بر اساس تغییر ناگهانی در اختلاف دو فاصله ادغام متوالی بود و طبق رابطه ۴ محاسبه می‌گردد:

رابطه (۴)
$$\Delta\alpha = \alpha_{i+1} - \alpha_i$$

که در آن $\alpha_i = 1,2,...,n-1$، iامین فاصله ادغام در دندروگرام حاصل از تجزیه خوشه‌ای بوده و n تعداد ژنوتیپ‌ها است. در نهایت برای تشخیص صحیح‌ترین گروه‌بندی حاصل از روش‌های مختلف تجزیه خوشه‌ای از روش تجزیه تابع تشخیص به روش خطی فیشر و برای تعیین انتساب اشتباه افراد در درون گروه‌ها نیز از روش u- اعتباری استفاده گردید (جابسن، ۱۹۹۲ و ولینگ، ۲۰۰۶). کلیه تجزیه‌های آماری شامل تجزیه خوشه‌ای و تجزیه تابع تشخیص با نرم افزار SPSS نسخه ۹ انجام شد.

نتایج و بحث

ابتدا روش‌های مختلف استاندارد کردن داده‌ها و داده‌های استاندارد نشده با شاخص‌های مختلف فاصله با استفاده از روش‌های وارد و متوسط فاصله بین دسته‌ها برای گروه‌بندی ژنوتیپ‌ها استفاده شد. سپس بهترین شاخص فاصله و روش استاندارد کردن داده‌ها در این سه روش انتخاب گردید و از آن‌ها برای مقایسه روش‌های مختلف تجزیه خوشه‌ای استفاده شد. تجزیه خوشه‌ای با استفاده از داده‌های حاصل از روش‌های مختلف استاندارد کردن و استاندارد نکردن داده‌ها با روش وارد و متوسط فاصله بین دسته‌ها نشان داد که شاخص‌های فاصله اقلیدسی و چبی‌چف ژنوتیپ‌ها را در سه گروه، شاخص‌های فاصله پیرسون و مینکوسکی، ژنوتیپ‌ها را در چهار گروه و شاخص‌های فاصله کوساین و بلوک آن‌ها را در پنج گروه قرار دادند. مقایسه گروه‌های حاصل با استفاده از تجزیه تابع تشخیص خطی فیشر حاکی از آن بود که شاخص فاصله اقلیدسی با استفاده از داده‌های

1. Minkowski
2. Chebychev
3. Cosine
4. Block

استاندارد شده به هر سه روش توانست ژنوتیپ‌ها را با احتمال صحت گروه‌بندی ۱۰۰ درصد بهتر از سایر شاخص‌های فاصله گروه‌بندی کرده و تفاوت بین ژنوتیپ‌ها را نشان دهد. احتمال انتساب اشتباه ژنوتیپ‌ها به گروهی که متعلق به آن نیستند در این روش صفر بود. شاخص فاصله بلوک نیز در مورد تمامی انواع داده‌ها کمترین احتمال ممکن را در صحیح نسبت دادن ژنوتیپ‌ها به گروه‌های مختلف دارا بود و با احتمال ۷۷/۸ درصد ضعیف‌ترین گروه‌بندی را در این روش‌های تجزیه خوشه‌ای ارائه داد (جدول ۲). این نتیجه علاوه بر تایید شاخص فاصله اقلیدسی به عنوان یک شاخص فاصله مناسب برای داده‌های مورفولوژیک و کمی، نشان داد که هر سه روش استاندارد کردن داده‌ها نتایج مشابهی ارائه داده و بهتر از استاندارد نکردن داده‌ها بودند.

سپس با توجه به نتایج بهتر شاخص فاصله اقلیدسی نسبت به سایر شاخص‌ها و همچنین یکسان بودن نتایج روش‌های استاندارد کردن داده‌ها در تجزیه خوشه‌ای در این دو روش (جدول ۲)، گروه‌بندی ژنوتیپ‌ها به روش‌های مختلف تجزیه خوشه‌ای با استفاده از شاخص فاصله اقلیدسی و استاندارد کردن داده‌ها به روش دوم (رابطه ۲) انجام شد و صحت گروه‌بندی‌ها در روش‌های مختلف به کمک تجزیه تابع تشخیص مورد تجزیه و تحلیل قرار گرفت و روش‌های مختلف تجزیه خوشه‌ای با هم مقایسه شدند.

جدول ۱: اسامی لاین‌ها، هیبریدها و توده‌های آزاد گرده‌افشان مورد مطالعه

ردیف	ژنوتیپ	شماره ژنوتیپ	نوع ژنوتیپ
۱	L 105	۱	لاین
۲	S 61	۲	لاین
۳	MO 17	۳	لاین
۴	K 1264-1	۴	لاین
۵	K 722	۵	لاین
۶	TVA 926	۶	لاین
۷	B 73	۷	لاین
۸	SC 108	۸	هیبرید سینگل کراس
۹	SC 301	۹	هیبرید سینگل کراس
۱۰	SC 604	۱۰	هیبرید سینگل کراس
۱۱	SC 647	۱۱	هیبرید سینگل کراس
۱۲	SC 704	۱۳	هیبرید سینگل کراس
۱۳	SC 711	۱۴	هیبرید سینگل کراس
۱۴	TWC 647	۱۲	هیبرید تری وی کراس
۱۵	SC 301	۱۵	توده حاصل از گرده‌افشانی آزاد
۱۶	SC 604	۱۶	توده حاصل از گرده‌افشانی آزاد
۱۷	SC 647	۱۷	توده حاصل از گرده‌افشانی آزاد
۱۸	SC 704	۱۸	توده حاصل از گرده‌افشانی آزاد

دندروگرام حاصل از تجزیه خوشه‌ای برای روش‌های مختلف در شکل‌های ۱ تا ۷ نشان داده شده است. نقطه برش بر اساس تغییر ناگهانی در اختلاف دو فاصله ادغام متوالی به‌دست آمد و دندروگرام‌های حاصل در آن نقاط برش داده شدند. بر اساس برش‌های ایجاد شده در دندروگرام‌ها، این روش‌ها در ۳ دسته کلی قرار گرفتند که روش‌های هر دسته، گروه‌بندی مشابهی داشتند. دسته اول شامل روش‌های متوسط فاصله بین دسته‌ها، دورترین همسایه‌ها و حداقل واریانس وارد بود. این سه روش مربوط به دسته اول، ژنوتیپ‌ها را در سه گروه قرار دادند و نتایج مشابهی داشتند (شکل‌های ۱، ۲ و ۳). دسته دوم تنها شامل روش متوسط فاصله بین و درون دسته‌ها بود و ژنوتیپ‌ها را در ۳ گروه دسته‌بندی کرد (شکل ۷)، اما گروه‌بندی حاصل با روش‌های دسته اول متفاوت بود و دسته سوم شامل روش نزدیک‌ترین همسایه‌ها، مرکزی و روش میانه‌ای بود و روش‌های این دسته ژنوتیپ‌ها را در ۴ گروه قرار دادند (شکل‌های ۴، ۵ و ۶).

ارزیابی گروه‌بندی حاصل از این روش‌ها، با توجه به دندروگرام‌ها نشان داد که در اکثر روش‌ها گروه‌بندی مطلوبی انجام گرفته است و حالت زنجیره‌ای مشاهده نمی‌شود، اما برای تشخیص این‌که کدامیک از روش‌ها گروه‌بندی مناسبتری انجام داده است از تجزیه تابع تشخیص استفاده شد. بدین منظور روش‌های مختلف تجزیه خوشه‌ای که در ۳ دسته با گروه‌بندی‌های متفاوت قرار گرفته بودند به‌وسیله تجزیه تابع تشخیص مورد ارزیابی قرار گرفته و صحت گروه‌بندی آن‌ها مورد بررسی قرار گرفت. توابع تشخیص به‌دست آمده برای روش‌های دسته اول که ژنوتیپ‌ها را در سه گروه قرار دادند و شامل روش‌های متوسط فاصله بین کلاسترها، حداقل واریانس وارد و دورترین همسایه‌ها بودند، به شرح ذیل بود:

رابطه (۵)

$$B_1 = -3594.7 + 19.22X_1 + 49.35X_5 - 458.57X_6 - 40.54X_{13} - 26.67X_{18} + 52.93X_{20}$$

رابطه (۶)

$$B_2 = -2462.37 + 16.39X_1 + 40.91X_5 - 381.04X_6 - 34.3X_{13} - 26.61X_{18} + 45.61X_{20}$$

رابطه (۷)

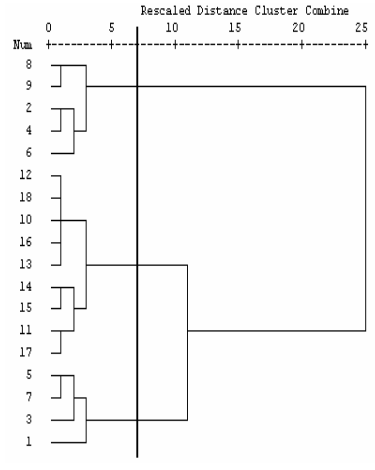
$$B_3 = -4546.38 + 25.3X_1 + 54.5X_5 - 527.18X_6 - 53.49X_{13} - 59.57X_{18} + 71.3X_{20}$$

که در آن X_1 : ارتفاع بوته، X_5 : طول برگ، X_6 : عرض برگ، X_{13} : تعداد ردیف دانه، X_{18} : تعداد روز تا ظهور کاکل و X_{20} : تعداد روز تا رسیدگی فیزیولوژیکی است و سایر متغییرها هم به دلیل معنی‌دار نبودن ضرایب آن‌ها در معادله وارد نشدند.

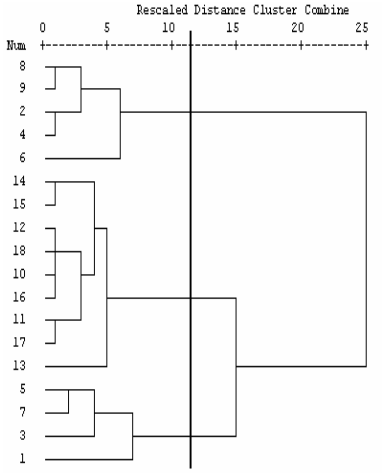
جدول ۲: ارزیابی صحت گروه‌بندی شاخص‌های مختلف فاصله، روش‌های مختلف استاندارد کردن دادها با تجزیه تابع تشخیص

روش تجزیه خوشه‌ای	روش استاندارد کردن	شاخص‌های فاصله							
		اقلیدسی	فاصله	چبی چف	فاصله	پیرسون	فاصله	مینکوسکی	فاصله	کوساین	فاصله	بلوک
متوسط فاصله بین دسته‌ها	استاندارد نکردن	۸۳/۳		۸۳/۳		۸۳/۳		۸۳/۳		۸۳/۳		۷۷/۸
	استاندارد با رابطه ۱	۱۰۰		۸۸/۹		۸۸/۹		۹۴/۴		۹۴/۴		۸۸/۹
	استاندارد با رابطه ۲	۱۰۰		۱۰۰		۱۰۰		۹۴/۴		۹۴/۴		۹۴/۴
	استاندارد با رابطه ۳	۱۰۰		۱۰۰		۹۴/۴		۹۴/۴		۸۸/۹		۹۴/۴
حداقل واریانس وارد	استاندارد نکردن	۸۳/۳		۸۳/۳		۸۳/۳		۱۰۰		۸۳/۳		۱۰۰
	استاندارد با رابطه ۱	۱۰۰		۱۰۰		۸۸/۹		۱۰۰		۹۴/۴		۱۰۰
	استاندارد با رابطه ۲	۱۰۰		۱۰۰		۱۰۰		۱۰۰		۹۴/۴		۱۰۰
	استاندارد با رابطه ۳	۱۰۰		۱۰۰		۸۸/۹		۱۰۰		۹۴/۴		۱۰۰

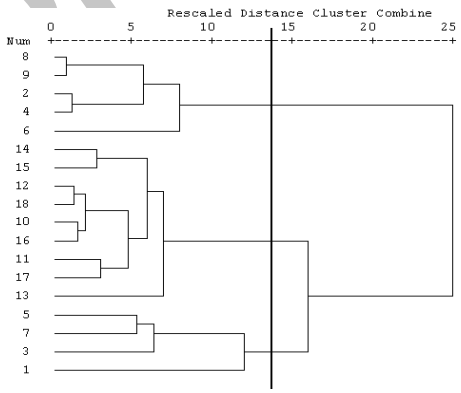
این توابع درصد بالایی از واریانس کل را توضیح می‌دهند و با استفاده از این توابع می‌توان ارقام جدید را به گروه‌های مربوطه منتسب نمود. ضرایب استاندارد شده صفات در توابع تشخیص نیز حاکی از این است که می‌توان مقدار عددی را برای هر ژنوتیپ با توجه به صفات مربوطه به دست آورد. از مقایسه این مقدار با مقادیر ارائه شده برای هر گروه و تابع تشخیص، می‌توان ژنوتیپ مذکور را به گروهی منتسب نمود که کم‌ترین فاصله را با آن داشته باشد. هم‌چنین با توجه به ضرایب صفات در هر تابع می‌توان به اهمیت نسبی هر صفت در تمایز بین گروه‌ها پی برد. به عنوان مثال عرض برگ در این سه تابع بیشترین تاثیر را دارد.



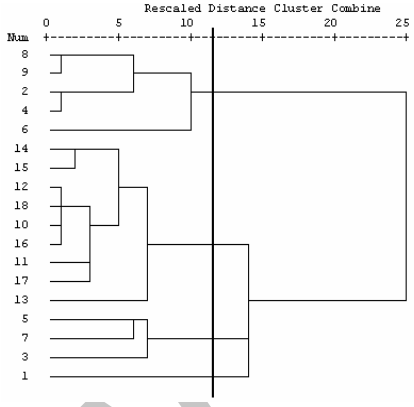
شکل ۱: دندروگرام روش حداقل واریانس وارد



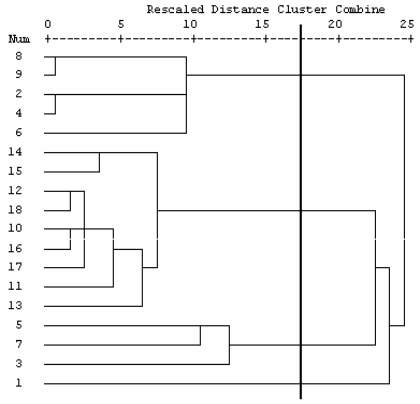
شکل ۲: دندروگرام روش دورترین همسایه‌ها



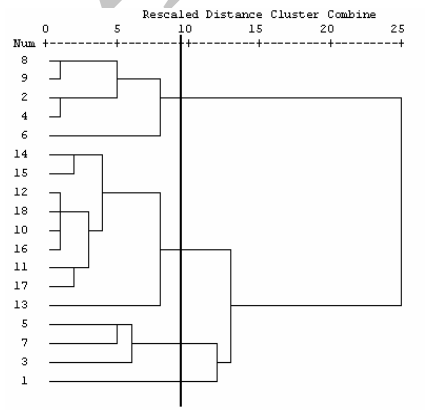
شکل ۳: دندروگرام روش متوسط فاصله بین کلاسترها



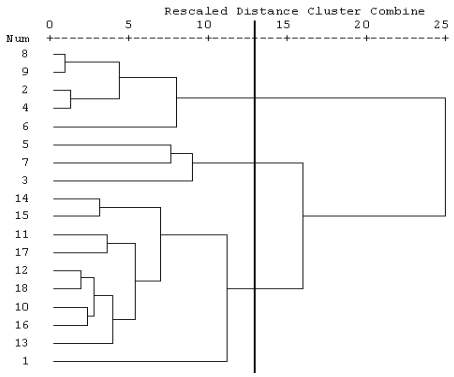
شکل ۴: دندروگرام روش مرکزی



شکل ۵: دندروگرام روش نزدیکترین همسایه



شکل ۶: دندروگرام روش میانه‌ای



شکل ۷: دندروگرام روش متوسط فاصله بین و درون کلاسترها

تجزیه تابع تشخیص خطی فیشر برای آزمون صحت گروه‌بندی اولیه ژنوتیپ‌ها در این دسته با استفاده از روش u- اعتباری نشان داد که هر سه گروه با احتمال ۱۰۰٪ صحیح گروه‌بندی شدند (جدول ۳).

توابع تشخیص به‌دست آمده برای روش‌های دسته دوم که شامل روش متوسط فاصله بین و درون کلاسترها بود و ژنوتیپ‌ها را در سه گروه قرار داد، به شرح ذیل بود:

رابطه (۸)

$$B_1 = -674.55 + 0.8X_1 + 16.69X_3 + 12.97X_{22}$$

رابطه (۹)

$$B_2 = -443.94 + 0.65X_1 + 11.59X_3 + 10.84X_{22}$$

رابطه (۱۰)

$$B_3 = -703.08 + 0.57X_1 + 20.25X_3 + 13.37X_{22}$$

که در آن X_1 : ارتفاع بوته، X_3 : تعداد گره و X_{22} : تعداد روز تا خاتمه گرده افشانی است و سایر متغییرها هم به دلیل معنی‌دار نبودن ضرایب آن‌ها در معادله وارد نشدند.

نتایج نشان داد که از بین سه گروه ایجاد شده یک گروه با احتمال ۹۰٪ صحیح و دو گروه دیگر کاملاً درست گروه‌بندی شدند. احتمال انتساب اشتباه برای این گروه‌بندی‌ها ۵/۶٪ به دست آمد به طوری که ژنوتیپ شماره یک (MO 17) بجای قرار گرفتن در گروه اول باید در گروه سوم قرار می‌گرفت (جدول۴). گروه‌های به دست آمده با روش نزدیک‌ترین همسایه‌ها، مرکزی و روش میانه‌ای نیز با روش u- اعتباری مورد ارزیابی قرار گرفتند و نتایج به‌دست آمده نشان داد که از بین چهار گروه ایجاد شده، یک گروه ۸۸/۹٪ صحیح، یک گروه کاملاً نادرست و دو گروه دیگر کاملاً درست گروه‌بندی شدند. درکل روش‌های این دسته توانسته‌اند با احتمال ۸۸/۹٪ ژنوتیپ‌ها را صحیح گروه‌بندی نمایند و احتمال انتساب اشتباه ژنوتیپ‌ها در این گروه ۱۱/۱۱٪ برآورد شد (جدول ۳). هم‌چنین نتایج نشان داد که ژنوتیپ‌ها بجای ۴ گروه باید در ۳ گروه قرار می‌گرفتند (جدول ۴). توابع تشخیص برآورد شده برای این روش هم به‌صورت زیر بود:

رابطه (۱۱)

$$B_1 = -135.17 + 21.O2X_8 + 6.05X_{12} - 0.35X_{15}$$

رابطه (۱۲)

$$B_2 = -118.79 + 14.23X_8 + 8.77X_{12} - 0.23X_{15}$$

رابطه (۱۳)

$$B_3 = -255.79 + 26.85X_8 + 10.75X_{12} - 0.55X_{15}$$

رابطه (۱۴)

$$B_4 = -201.56 + 22.33X_8 + 8.35X_{12} - 0.24X_{15}$$

که در آن X_8 : تعداد کل برگ، X_{12} : تعداد ردیف دانه و X_{15} : وزن دانه است و سایر متغییرها هم به دلیل معنی‌دار نبودن ضرایب آن‌ها با توجه به روش گام به گام مورد استفاده در تجزیه تابع تشخیص در معادله وارد نشدند.

در مقایسه روش‌های مختلف تجزیه خوشه‌ای، نوع داده‌های مورد استفاده و تکنیک‌های مختلف تجزیه خوشه‌ای برای گروه‌بندی ژنوتیپ‌ها می‌توان گفت که فاصله اقلیدسی (جدول ۲) از سایر شاخص‌های فاصله بهتر بود، چرا که از مزایای این فاصله این است که اگر ماتریس داده‌ها دارای مقادیر از دست رفته باشد، باز هم می‌توان از این شاخص استفاده نمود. هم‌چنین از این شاخص به نحو مطلوب می‌توان برای ارزیابی فاصله بین ژنوتیپ‌ها به ویژه برای صفات مورفولوژیک و کمی استفاده نمود (جابسن، ۱۹۹۲). این نتیجه با انجام تجزیه تابع تشخیص در این تحقیق نیز به دست آمد، به‌طوری که در تمامی روش‌های مورد استفاده، احتمال گروه‌بندی صحیح ژنوتیپ‌ها با این شاخص فاصله از ۸۳/۳٪ در روش میانه‌ای تا ۱۰۰٪ در روش وارد و متوسط فاصله بین کلاسترها متغیر بود.

جدول ۳: جدول نسبت موفقیت افراد درون گروه‌ها با تابع تشخیص

روش‌های گروه‌بندی	تعداد افراد در هر گروه	گروه‌ها	تعداد افراد پیش‌بینی شده در هر گروه با تابع تشخیص		
			۱	۲	۳
دسته اول	۴	۱	۴ ٪۱۰۰	۰	۰
	۵	۲	۰	۵ ٪۱۰۰	۰
	۹	۳	۰	۰	۹ ٪۱۰۰
	تعداد افراد در هر گروه	گروه‌ها	۱	۲	۳
دسته دوم	۱۰	۱	۹ ٪۹۰	۰	۱ ٪۱۰
	۵	۲	۰	۵ ٪۱۰۰	۰
	۳	۳	۰	۰	۳ ٪۱۰۰
	تعداد افراد در هر گروه	گروه‌ها	۱	۲	۳	۴
دسته سوم	۱	۱	۰	۰	۰	۱ ٪۱۰۰
	۳	۲	۰	۳ ٪۱۰۰	۰	۰
	۵	۳	۰	۰	۵ ٪۱۰۰	۰
	۹	۴	۰	۱ ٪۱۱/۱	۰	۸ ٪۸۸/۹

علاوه بر شاخص فاصله، نوع داده‌های مورد استفاده در تجزیه خوشه‌ای و گروه‌بندی ژنوتیپ‌ها، یعنی به کار بردن داده‌های استاندارد شده و یا استفاده از داده‌های خام اولیه، می‌تواند مطلوبیت گروه‌بندی حاصل را تحت تاثیر قرار دهد. بدیهی است که برای انجام تجزیه خوشه‌ای، صفات مختلفی مورد اندازه‌گیری قرار می‌گیرند که دارای مقیاس اندازه‌گیری متفاوتی هستند. علاوه بر آن، تنوع یا واریانس این صفات نیز بسیار متنوع است، به‌طوری که اگر مستقیماً از داده‌های خام اولیه برای انجام تجزیه خوشه‌ای استفاده گردد، ش فاصله مورد استفاده برای ارزیابی میزان فاصله ژنتیکی بین ژنوتیپ‌ها و به دنبال آن گروه‌بندی انجام شده کاملاً تحت تاثیر صفات با مقیاس کوچک‌تر (که دارای اعداد بزرگ‌تری هستند) و یا صفات با واریانس بزرگتر قرار گرفته و لذا عموماً نتایج مطلوبی به دست نخواهد آمد (جابسن، ۱۹۹۲ و رومسبرگ، ۱۹۹۰). با استاندارد کردن داده‌ها، اثر صفات با مقیاس کوچک‌تر و واریانس بزرگ‌تر تعدیل شده و همه صفات دارای وزن یکسانی خواهند شد و لذا در تجزیه خوشه‌ای و تشکیل گروه‌های حاصل سهم یکسانی نیز خواهند داشت. نتایج حاصل از تجزیه تابع تشخیص و برآورد احتمال انتساب صحیح ژنوتیپ‌ها به گروه‌ها نیز این موضوع را مورد تایید قرار داد، به‌طوری که گروه بندی حاصل با داده‌های استاندارد شده در تمامی روش‌های تجزیه خوشه‌ای کاملاً متفاوت از گروه-های حاصل از داده‌های استاندارد نشده بود (جدول ۲). اما تفاوتی بین نتایج حاصل از داده‌های استاندارد شده در سه روش (روابط ۱، ۲ و۳) وجود نداشت و تجزیه تابع تشخیص، گروه‌بندی حاصل از هر ۳ مجموعه داده‌های

استاندارد شده را کاملاً مشابه و صحت آن‌ها را از ۷۲/۲٪ در روش میانه‌ای با شاخص فاصله مینکوسکی تا ۱۰۰٪ گروه‌بندی صحیح در روش‌های متوسط فاصله بین کلاسترها، دورترین همسایه‌ها و حداقل واریانس وارد با شاخص فاصله اقلیدسی محاسبه نمود (جدول ۲).

جدول ۴: پیش‌بینی وضعیت ژنوتیپ‌ها در گروه‌ها بر اساس تجزیه تابع تشخیص

تیمارها	گروه‌های واقعی دسته اول	گروه‌های پیش بینی شده برای دسته اول	گروه‌های واقعی دسته دوم	گروه‌های پیش بینی شده برای دسته دوم	گروه‌های واقعی دسته سوم	گروه‌های پیش بینی شده برای دسته سوم
۱	۱	۱	۱	۳**	۱	۴**
۲	۲	۲	۲	۲	۲	۲
۳	۱	۱	۳	۳	۳	۳
۴	۲	۲	۲	۲	۲	۲
۵	۱	۱	۳	۳	۳	۳
۶	۲	۲	۲	۲	۲	۲
۷	۱	۱	۳	۳	۳	۳
۸	۲	۲	۲	۲	۲	۲
۹	۲	۲	۲	۲	۲	۲
۱۰	۳	۳	۱	۱	۴	۴
۱۱	۳	۳	۱	۱	۴	۴
۱۲	۳	۳	۱	۱	۴	۴
۱۳	۳	۳	۱	۱	۴	۴
۱۴	۳	۳	۱	۱	۴	۲**
۱۵	۳	۳	۱	۱	۴	۴
۱۶	۳	۳	۱	۱	۴	۴
۱۷	۳	۳	۱	۱	۴	۴
۱۸	۳	۳	۱	۱	۴	۴

** گروه‌بندی اشتباه

روش تجزیه خوشه‌ای مورد استفاده نیز مستقیماً می‌تواند گروه‌بندی‌های حاصل را تحت تاثیر قرار دهد. مطلوب بودن روش‌های متوسط فاصله بین کلاسترها و حداقل واریانس وارد با تجزیه تابع تشخیص نیز به اثبات رسید به‌طوری‌که احتمال انتساب صحیح ژنوتیپ‌ها به گروه‌های مربوطه ۱۰۰٪ برآورد شد. علاوه بر آن، دندروگرام حاصل از هر دو روش فوق، گروه‌بندی بسیار مطلوبی داشته و با خصوصیات مورفولوژیک و کمی ژنوتیپ‌های مورد مطالعه مطابقت داشت. روش‌های متوسط فاصله بین کلاسترها و حداقل واریانس وارد نتیجه بهتری از سایر روش‌ها ارائه می‌دهند (جابسن، ۱۹۹۲). روش متوسط فاصله بین کلاسترها از متوسط فاصله بین ژنوتیپ‌ها یا گروه‌ها استفاده کرده و دو گروهی که متوسط فاصله بین آن‌ها کم‌ترین مقدار را داشته باشد در هم ادغام می‌شوند. در مقابل، روش حداقل واریانس وارد بر مبنای محاسبه مجموع مربعات درون گروهی استوار بوده و لذا فاصله بین گروه‌ها یا ژنوتیپ‌ها را بزرگ‌تر کرده و بهتر نشان می‌دهد. به این دلیل به نظر می‌رسد این روش خصوصاً برای صفات کمی و مزرعه‌ای گروه‌بندی بهتری ایجاد نماید، چرا که در این روش ترکیب‌های دوبه‌دوی بین گروه‌ها در نظر گرفته شده و ترکیبی که مجموع مربعات درون گروهی آن کم‌ترین مقدار باشد به عنوان بهترین ترکیب انتخاب و گروه‌بندی مربوطه ایجاد می‌شود. به این ترتیب یک نوع معیار بهینگی همانند روش‌های تجزیه برآورد می‌شود، اگرچه این معیار بهینگی مطلق نیست. محققین دیگر نیز با انجام تجزیه تابع تشخیص صحت گروه‌بندی تجزیه خوشه‌ای را بررسی کردند، از جمله موردا و همکاران (۲۰۰۳) با تجزیه تابع تشخیص نشان دادند که گروه‌بندی به روش حداقل واریانس وارد و شاخص فاصله

اقلیدسی با احتمال ۹۴/۴٪ بهتر از روش‌های دیگر بوده است. هم‌چنین جینز و همکاران (۲۰۰۳) و بالوچی و همکاران (۲۰۰۱) نیز نشان دادند که روش‌های حداقل واریانس وارد و UPGMA گروه‌بندی بهتری انجام داده‌اند. که نتایج آن‌ها با این پژوهش در یک راستا بود و حاکی از گروه‌بندی مطلوب‌تر این روش‌ها در نشان دادن شباهت‌ها و تفاوت‌های بین ژنوتیپ‌ها و تیمارهای مورد مطالعه است.

نتیجه‌گیری

با توجه به نتایج حاصل از این آزمایش می‌توان این‌گونه نتیجه گرفت که انتخاب روش صحیح تجزیه خوشه‌ای و شاخص فاصله مورد استفاده در بررسی تنوع ژنتیکی و گروه‌بندی ژنوتیپ‌ها از اهمیت زیادی برخوردار است، به طوری که یک انتخاب نادرست می‌تواند ژنوتیپ‌هایی را در یک گروه قرار دهد که شباهت زیادی با یکدیگر نداشته باشند. اگرچه تجزیه و تحلیل شاخص‌های مختلف فاصله و روش‌های متفاوت تجزیه خوشه‌ای در این مطالعه نشان داد که استفاده از شاخص فاصله اقلیدسی بر اساس داده‌های استاندارد شده و انجام تجزیه خوشه‌ای با روش‌های حداقل واریانس "وارد" و متوسط فاصله بین کلاسترها گروه‌بندی بهتری از ژنوتیپ‌های ذرت انجام می‌دهند، اما توصیه می‌شود که در هر نوع مطالعه‌ای، روش‌های مختلف تجزیه خوشه‌ای با استفاده از تجزیه تابع تشخیص مورد آزمون قرار گیرند و بهترین روش در گروه‌بندی ژنوتیپ‌ها انتخاب گردند.

سپاسگزاری

بدین‌وسیله از ریاست محترم مرکز تحقیقات کشاورزی و منابع طبیعی همدان و مسول محترم بخش اصلاح بذر این مرکز بخاطر در اختیار قرار دادن امکانات لازم جهت انجام این تحقیق، کمال تقدیر و تشکر خود را اغلام می‌دارم. همچنین از آقای مهندس بانکه‌ساز به سبب راهنمایی‌های ارزنده‌شان کمال تشکر را دارم.

منابع

Balocchi, L. O., Caballero, J. V. and Smith, R. R. 2001. Characterization and agronomic variability of 125 ecotypes of *Bromus valdivianus* Phil, collected from Valdivia province. Agro. Sur. 29 (1): 64-77.

Bretting, P. K., Goodman M. M. and Studer, C. W. 1990. Isozymatic variation in Guatemalan races of maize. American Journal of Botany 77: 211-225.

Crossa, J., Basford, K., Taba, S., De Lacy, I. and Silva, E. 1995. Three mode analyses of maize using morphological and agronomic attributes measured in multilocation trials. Crop Science 35: 1483-1491.

Fernandez, G. C. J. 2006. Discriminant analysis: A powerful classification technique in data mining; http://www2.sas. com/proceedings/sugi 27/p247-27. pdf.

Jaynes, D. B., Kaspar, T. C., Colvin, T. S. and James, D. E. 2003. Cluster analysis of spatiotemporal corn yield patterns in an Iowa field, Agron. J. 95 (3): 574-586.

Jiang, J. H., Tsenkova, R. and Ozaki, Y. 2001. Principle discriminant variate method for classification of multicollinear data: principle and applications. Analytical Sciences. 17: 471-474.

Jobson, J. D. 1992. Applied multivariate data analysis. Volum Π. Categorical and multivariate methods. New york Springer-Verlag.

Mendez, M. A., Hodar, C., Vulpe, C., Gonzalez, M. and Cambiazo, V. 2002. Discriminant analysis to evaluate clustering of gene expression data. Federation of European Biochemical Societies. 522: 24-28.

Moreda, A. P., Fisher, A. and Hill, S. J. 2003. The classification of tea according to region of origin using pattern recognition techniques and trace metal data. Journal of Food Composition and Analysis. 16 (2): 195-211.

Romesburg, H. C. 1990. Cluster analysis for researchers, Krieger Publishing, Malabar. Florida.

Welling, M. 2006. Fisher linear discriminant analysis. http://www.ics.uci.edu /~welling/classnotes/papers_class/Fisher-LDA.pdf.

The Evaluation of Grouping Accuracy of Different Cluster Analysis Methods

Ramezani[1*], M., Rahimi[1], M., Samezade Lahije [2], H. A. and Rahimi[3], F.

Abstract

In this study, clusters validity of several methods of cluster analysis based on different methods of data standardization and different distance criteria will be evaluated with discrimination function analysis. To reach the aims, 7 inbred lines and 7 hybrids progenies of crossing between them and 4 open pollinated F_1 hybrids were grown in an unbalanced grouped design with 3 replications and 22 agronomic characteristics were evaluated on genotypes. The discrimination function analysis showed that the Euclidean distance criterion was better than other distances criteria and a desirable clustering got based on it. Also, data standardization methods had same clustering and were better than data un-standardization. The dendrograms observation of several cluster analysis methods showed that the Unweighted Pair Group Method with Arithmetic (UPGMA), Complete linkage and Ward's minimum variance methods were better than the other methods. Also, the Fisher's linear discrimination analysis was fulfilled and it was observed that UPGMA, Complete linkage and Ward's minimum variance methods with performing valid clustering about 100% better than other methods could use of initial information of data and classified genotypes.

Keywords: Cluster analysis, discrimination function analysis, *Zea mays* L.

1 And 2. M.Sc. Students and Assistant Professor respectively, Department. of Agronomy and Plant Breeding, Faculty of Agricultural Sciences, Guilan University, Rasht

3. M.Sc. Student of Environmental Planning and Management, Faculty of Environment, Tehran University, Tehran

∗. *Corresponding Author*