

ارائه یک الگوریتم انتخاب مشخصه بهینه بدون نظارت

حمیدرضا کاکائی مطلق^۱، مهدی ملازاده گل محله^{۲*}، بابک تیمورپور^۳

۱- مربی، دانشگاه جامع امام حسین^(ع)

۲- مربی، دانشگاه جامع امام حسین^(ع)

۳- استادیار، دانشگاه تربیت مدرس

(دریافت: ۹۲/۰۵/۹؛ پذیرش: ۹۴/۱۰/۲۲)

چکیده

انتخاب بردار مشخصه مناسب برای حداکثر نمودن موفقیت یک ماشین دسته‌بندی‌کننده بسیار موثر است. در این مقاله با استفاده از ترکیب روش‌های مختلف محاسبه تابع هسته، یک الگوریتم انتخاب مشخصه بهینه بدون نظارت پیشنهاد گردیده است. بردار مشخصه به دست آمده از الگوریتم پیشنهادی، صحت خروجی دسته‌بندی‌کننده شبکه عصبی پس‌انتشار خطا را حداکثر می‌گرداند. در این مقاله برای مطالعه موردی از دسته‌بندی استاندارد تصاویر فشرده شده مبتنی بر کدگذاری تبدیلی و تصاویر فشرده نشده با استفاده از رشته‌بیت آن‌ها استفاده می‌گردد. استانداردهای مورد نظر برای دسته‌بندی، استانداردهای JPEG و JPEG2000 و تصاویر فشرده نشده با فرمت TIFF می‌باشند. با استفاده از بردار مشخصه به دست آمده از الگوریتم پیشنهادی، صحت دسته‌بندی‌کننده در حدود ۹۸٪ می‌گردد.

واژه‌های کلیدی: بردار مشخصه، انتخاب بردارهای مشخصه، شبکه عصبی، دسته‌بندی، استاندارد فشرده‌سازی تصویر

داده دیجیتال تصویری با استفاده از رشته‌بیت آن می‌باشد. با توجه به این که تصاویر اغلب با استفاده از الگوریتم‌هایی همچون JPEG و JPEG2000 فشرده می‌شوند، لذا در این مقاله رشته‌بیت این دو استاندارد برای دسته‌بندی استفاده می‌شوند. همچنین فرمت TIF نیز به عنوان خروجی فشرده نشده تصویر و یک کلاس مستقل در نظر گرفته شده است. در ادامه مقاله ابتدا به معرفی الگوریتم انتخاب بردار مشخصه مناسب پرداخته شده است. سپس چهارچوب کلی ارزیابی الگوریتم پیشنهادی بر مبنای دسته‌بندی رشته‌بیت تصاویر بیان گردیده است. استخراج مشخصه از رشته‌بیت تصاویر فشرده شده و فشرده نشده توضیح داده شده است. سپس با استفاده از شبکه عصبی دسته‌بندی انجام گرفته است. در بخش پایانی مقاله نتایج شبیه‌سازی و نتیجه‌گیری آورده شده است.

۲- الگوریتم انتخاب بردار مشخصه مناسب

اگر تعداد کل مشخصه‌ها برابر N باشد، تعداد کل زیرمجموعه‌های ممکن برابر 2^N می‌شود. این تعداد برای N های متوسط هم خیلی زیاد است. روش‌های مختلف انتخاب مشخصه، تلاش می‌کنند تا از میان 2^N زیرمجموعه‌ها کاندید، بهترین زیرمجموعه را پیدا کنند. در تمام این روش‌ها بر اساس کاربرد و نوع تعریف، زیرمجموعه‌ای به عنوان جواب انتخاب می‌شود، که بتواند مقدار یک تابع ارزیابی را بهینه کند. با وجود این که هر

۱- مقدمه

در سیستم‌های ارتباطی دیجیتال پیشرفته، توانایی دسته‌بندی^۱ و دسته‌بندی کردن داده‌ها برای بسیاری از سازمان‌دهی‌ها^۲ مطلوب می‌باشد. به عبارتی دسته‌بندی و تفکیک خودکار داده‌های دیجیتال جمع‌آوری شده و ذخیره کردن آن‌ها به منظور تحلیل بیشتر امری ضروری است.

مساله انتخاب مشخصه، یکی از مسائلی است که در مبحث یادگیری ماشین و همچنین شناسایی آماری الگو مطرح است. این مساله در بسیاری از کاربردها (مانند طبقه‌بندی) اهمیت به‌سزائی دارد، زیرا در این کاربردها تعداد زیادی مشخصه وجود دارد، که بسیاری از آنها یا بلااستفاده هستند و یا این که بار اطلاعاتی چندانی ندارند. حذف نکردن این مشخصه‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد مورد نظر بالا می‌برد. و علاوه بر این باعث می‌شود که اطلاعات غیر مفید زیادی را به همراه داده‌های مفید ذخیره کنیم.

داده دیجیتال می‌تواند مربوط به متن، صدا، تصویر و یا ترکیبی از این قبیل باشد. در این بررسی هدف ما دسته‌بندی

نحوه محاسبه ماتریس شباهت در خروجی الگوریتم تاثیرگذار است. در این مقاله تلاش نموده‌ایم تا روشی برای محاسبه ماتریس شباهت ارائه نماییم که خروجی دسته‌بندی کننده دارای دقت دسته‌بندی بالاتری باشد.

برای به‌دست آوردن ماتریس شباهت بهینه از دو روش محاسبه ماتریس شباهت متفاوت زیر که در [۶] و بیان شده است، استفاده نمودیم. در روش اول که در رابطه (۱) آمده است، میزان شباهت را بر مبنای محاسبه فاصله بین دو بردار به‌دست می‌آورد.

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\delta^2}} \quad (1)$$

در روش دوم که در رابطه (۲) آمده است، میزان شباهت را بر مبنای محاسبه زاویه دو بردار به‌دست می‌آورد.

$$S_{ij} = (x_i \cdot x_j + 1)^\delta \quad (2)$$

در روش پیشنهادی به‌منظور محاسبه دقیق‌تر میزان شباهت، از ترکیب هر دو معیار محاسبه شباهت استفاده شده است. پس با ترکیب نتیجه به‌دست‌آمده از روش‌های بالا ماتریس شباهت نهایی را محاسبه می‌نماییم. برای ترکیب ماتریس‌های شباهت محاسبه شده، ابتدا آنها را به روش گفته شده در بخش (۴-۳) نرمالیزه نمودیم و سپس از ترکیب خطی آنها (به‌عنوان ساده‌ترین حالت ممکن) به‌صورت زیر استفاده نموده‌ایم. البته می‌توان ترکیب‌های مختلفی را در نظر گرفت که متناسب با درجه آن ترکیب، پیچیدگی محاسبه ترکیب بهینه افزایش خواهد یافت.

$$S_{ij} = \alpha S_{ij}^1 + \beta S_{ij}^2 \quad (3)$$

برای محاسبه دقیق و بهینه متغیرهای آزاد α و β و δ می‌توان از الگوریتم‌های بهینه‌سازی مانند الگوریتم ژنتیک استفاده نمود. مقدار بهینه زمانی به‌دست می‌آید که برای داده اعتبارسنجی بهترین جواب را به‌دست آوریم.

۳- چهارچوب کلی ارزیابی الگوریتم پیشنهادی

بر مبنای دسته‌بندی رشته‌بیت تصاویر

به‌طور کلی، دسته‌بندی شامل مراحل استخراج مشخصه و انتخاب مشخصه‌های مناسب و دسته‌بندی می‌باشد [۲]. فرآیند دسته‌بندی رشته‌بیت تصاویر، مبتنی بر مقادیر بردار مشخصه است. هرچه بردار مشخصه بهتری به دسته‌بندی کننده داده شود خروجی آن نیز بهتر خواهد بود. لذا انتخاب بردار مشخصه مناسب در موفقیت دسته‌بندی بسیار موثر خواهد بود. در این مقاله برای یافتن مشخصه‌های مناسب استفاده از الگوریتم بهینه‌شده‌ای پیشنهاد گردیده است، که با کمک آن بردار مشخصه مناسب مشخص می‌گردد. دسته‌بندی خود شامل دو مرحله آموزش و

روشی سعی می‌کند که بتواند، بهترین مشخصه‌ها را انتخاب کند، اما با توجه به وسعت جواب‌های ممکن، و این‌که این مجموعه‌های جواب به‌صورت توانی از N افزایش پیدا می‌کنند، پیدا کردن جواب بهینه مشکل و در N های متوسط و بزرگ بسیار پرهزینه است.

به‌طور کلی روش‌های مختلف انتخاب مشخصه را بر اساس نوع جستجو به دسته‌های مختلفی تقسیم‌بندی می‌کنند. در بعضی روش‌ها تمام فضای ممکن جستجو می‌گردد. در سایر روش‌ها که می‌تواند مکاشفه‌ای و یا جستجوی تصادفی باشد، در ازای از دست دادن مقداری از کارآئی، فضای جستجو کوچکتر می‌شود.

۲-۱- الگوریتم انتخاب مشخصه طیفی

الگوریتم‌های انتخاب مشخصه ابعاد بردار مشخصه را به‌منظور بالا بردن کارآئی ماشین‌های یادگیری کاهش می‌دهند. الگوریتم‌های انتخاب مشخصه بر مبنای نوع یادگیری به دو دسته کلی با نظارت^۱ و بدون نظارت^۲ تقسیم‌بندی می‌گردند. الگوریتم استفاده‌شده در این مقاله الگوریتم انتخاب مشخصه طیفی [۶] است. این الگوریتم بر مبنای نظریه گراف طیفی و یک الگوریتم انتخاب مشخصه بدون نظارت می‌باشد.

Algorithm 1: SPEC

Input: $X, \gamma(\cdot), k, \hat{\varphi} \in \{\hat{\varphi}_1, \hat{\varphi}_2, \hat{\varphi}_3\}$
Output: SF_{SPEC} - the ranked feature list
 1 construct \mathbb{S} , the similarity set from X (and Y);
 2 construct graph G from \mathbb{S} ;
 3 build W, D and L from G ;
 4 for each feature vector f_i do
 5 $\hat{f}_i \leftarrow \frac{D^{\frac{1}{2}} f_i}{\|D^{\frac{1}{2}} f_i\|}; \quad SF_{SPEC}(i) \leftarrow \hat{\varphi}(F_i);$
 6 end
 7 ranking SF_{SPEC} in ascending order for $\hat{\varphi}_1$ and $\hat{\varphi}_2$, or descending order for $\hat{\varphi}_3$;
 8 return SF_{SPEC} ;

شکل (۱). الگوریتم انتخاب مشخصه طیفی

این الگوریتم با دریافت مجموعه‌ای از بردارهای مشخصه و ماتریس شباهت آن‌ها، برای هر مشخصه وزنی بین صفر تا ۱ مشخص می‌نماید. با کمک این وزن‌ها می‌توان مشخصه‌ها را اولویت‌بندی نمود و برای دسته‌بندی استفاده کرد.

۲-۲- بهینه‌سازی الگوریتم دسته‌بندی

الگوریتم انتخاب مشخصه طیفی برای انتخاب مشخصه‌های مناسب از ماتریس شباهت بردارهای مشخصه استفاده می‌نماید.

$$F_3 = \text{Skew} \frac{1}{N} \sum_{i=1}^N \left[\frac{x_i - F_1}{F_2} \right]^3 \quad (6)$$

• **درجهٔ اوج در یک نمودار آماری^۴ یا پخی:**

درجهٔ اوج در یک نمودار آماری، نقطهٔ پیک یا همواری توزیع را نسبت به توزیع نرمال نشان می‌دهد و طبق رابطهٔ زیر به دست می‌آید:

$$F_4 = \text{Kurt} = \frac{1}{N} \sum_{i=1}^N \left[\frac{x_i - F_1}{F_2} \right]^4 \quad (7)$$

• **خودهمبستگی^۵:**

همبستگی بین دو دادهٔ متفاوت ولی متجانس با قیاس مستقیم برهم منطبق شدهٔ هر دوی آن‌ها و با شیفت یافته به چپ یا راست یکی از آن‌ها، تعریف می‌شود [۳]. همبستگی گسستهٔ دو تابع نمونه برداری شدهٔ X_k و Y_k با N پریود به صورت زیر تعریف می‌شود:

$$F_5 = R_{X,Y}(n) = \frac{\sum_{k=0}^{N-1} X_{k+n} Y_k^*}{N} \quad (8)$$

خودهمبستگی گسستهٔ تابع نمونه برداری شدهٔ X_k عیناً برابر با همبستگی گسستهٔ تابع با خودش می‌باشد.

برای به دست آوردن مشخصه مناسب، چندین آزمایش در تأخیرهای زمانی متفاوت انجام شد و مقدار خودهمبستگی به ازای $n = 0$ انتخاب شد.

• **انرژی دنبالهٔ X:**

انرژی دنبالهٔ X عبارت است از:

$$F_6 = \text{Energy} = \frac{1}{N} \sum_{i=1}^N |x_i|^2 \quad (9)$$

• **تعداد تکرار رشته‌ای با طول مشخص:**

$$F_7 = \text{تعداد تکرار رشتهٔ تمام صفر به طول } 8 \quad (10)$$

$$F_8 = \text{تعداد تکرار رشتهٔ تمام یک به طول } 8 \quad (11)$$

۴-۲- **مشخصه‌های آماری استخراج شده از**

هیستوگرام رشته‌بیت تصاویر

هیستوگرام از شمارش تعداد سمبل‌های با مقادیر یکسان در دنبالهٔ X به دست می‌آید.

هیستوگرام برای رشته‌بیت تصویر فشرده شده و فشرده نشدهٔ Barbara با ابعاد 512×512 در شکل‌های ۲، ۳ و ۴ نشان داده شده است. در این شکل‌ها i در محور افقی بیانگر مقادیری است که بایت‌ها می‌توانند داشته باشند (در اینجا ۰ تا ۲۵۵) و $h(i)$ در محور عمودی بیانگر تعداد بایت با مقادیر یکسان می‌باشد.

تست می‌باشد. بردارهای مشخصه استخراج شده از رشته‌بیت تصاویر به دو مجموعهٔ آموزش و تست تقسیم می‌شوند. در مرحلهٔ آموزش یک مدل آموزش دیده توسط بردارهای مشخصه مجموعهٔ آموزش حاصل می‌شود و در مرحلهٔ تست از مدل آموزش دیده برای دسته‌بندی بردارهای مشخصه مجموعهٔ تست استفاده می‌شود.

۴- **استخراج مشخصه از رشته‌بیت تصاویر**

فشرده شده و فشرده نشده

اولین گام استخراج مشخصه‌هایی است که قرار است در دسته‌بندی از آن‌ها استفاده شود. هر چقدر مشخصه‌های استخراج شده بتوانند تمایز بین رشته‌بیت‌ها را بهتر نشان دهند به طبع عملیات دسته‌بندی با سهولت و کارایی بالاتری امکان پذیر است. بنابراین مشخصه‌های آماری مختلفی از رشته‌بیت و هیستوگرام به دست آمده از آن استخراج می‌شوند و بردار مشخصه را تشکیل می‌دهند.

۴-۱- **مشخصه‌های آماری استخراج شده از رشته‌بیت**

تصاویر

دنبالهٔ گسستهٔ $X = (x_1, x_2, \dots, x_N)$ را در نظر بگیرید (این دنباله از تبدیل رشته‌بیت تصاویر به N سمبل m بیتی به دست آمده است). مشخصه‌های مختلفی از رشته بیت می‌توان استخراج کرد [۵]. بعضی از مشخصه‌های مهم و پرکاربرد از این دنباله استخراج شده است:

• **میانگین^۱:**

میانگین دنبالهٔ X عبارت است از:

$$F_1 = \frac{1}{N} \sum_{i=1}^N x_i \quad (4)$$

• **انحراف معیار^۲:**

انحراف معیار دنبالهٔ X عبارت است از:

$$F_2 = \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - F_1)^2} \quad (5)$$

• **چولگی^۳ یا عدم تقارن:**

پارامتر چولگی، عدم تقارن توزیع پیرامون میانگین را مشخص می‌کند و طبق رابطهٔ زیر به دست می‌آید:

• آنترپی^۲:

اندازه گیری اطلاعات موجود در یک الگو را آنترپی می گویند. اگر $X = (x_1 x_2 \dots x_N)$ الگوی مورد نظر برای محاسبه آنترپی باشد و اگر $t_1 t_2 \dots t_M$ مقادیر مشخص در الگوی X با احتمال های $p_1 p_2 \dots p_M$ باشند ($M < N$) در این صورت آنترپی دنباله به صورت زیر محاسبه می شود:

$$F_{12} = H(X) = -\sum_{i=1}^M P(i) \cdot \log_2 P(i) \quad (15)$$

بنابراین هر الگویی به شکل دنباله X به بردار 12 بعدی $F = (F_1, \dots, F_{12})$ تبدیل می شود. این 12 مشخصه، که از دنباله X یا هیستوگرام آن استخراج می شوند، در مقادیر عددی اشان در بسیاری موارد فرق دارند. بنابراین همه آنها نیاز به نرمالیزه شدن در فاصله ای معین، یعنی صفر تا یک دارند. نرمال سازی مشخصه ها در ادامه شرح داده می شود.

۴-۳- نرمال سازی مشخصه ها

نرمالیزه کردن همه مقادیر مشخصه ها برای قرار گرفتن در بازه صفر و یک معمول می باشد. اگر n الگو هر یک دارای k بعد در کلاسی وجود داشته باشند، مقدار مشخصه نرمال شده $V(i, j)$ (به ازای $i = 1, \dots, n$ و $j = 1, \dots, k$) با محاسبه مقادیر $t_1 = \text{Min}(V(i, j))$ و $t_2 = \text{Max}(V(i, j))$ توسط رابطه زیر به دست می آید:

$$\text{Normal}V(i, j) = \frac{V(i, j) - t_1}{t_2 - t_1} \quad (16)$$

۵- دسته بندی با استفاده از شبکه عصبی

شبکه عصبی از بردار ورودی و هدف برای تقریب زدن یک تابع و یافتن رابطه بین ورودی و خروجی و دسته بندی ورودی ها استفاده می کند [۴].

برای دسته بندی ورودی ها از شبکه عصبی پس انتشار خطا^۳ با الگوریتم LM^۴ استفاده کرده ایم. این شبکه دارای یک لایه مخفی با تابع tan-sigmoid می باشد. الگوهای ورودی به این دسته بندی کننده، بردارهای مشخصه انتخاب شده هستند. تعداد نرون^۵ ورودی شبکه با بردار ورودی و تعداد نرون خروجی با خروجی مطلوب مطابقت می کند. از این رو این دو مورد ثابت هستند. تنها تغییر کننده تعداد نرون ها در لایه مخفی و همچنین توجه به اثر آن در یادگیری شبکه در حالت آموزش داده ها می باشد.

مینیمم و ماکزیمم هیستوگرام طبق رابطه های زیر محاسبه می شوند:

$$F_9 = \text{مقدار مینیمم هیستوگرام} = \min(h) \quad (12)$$

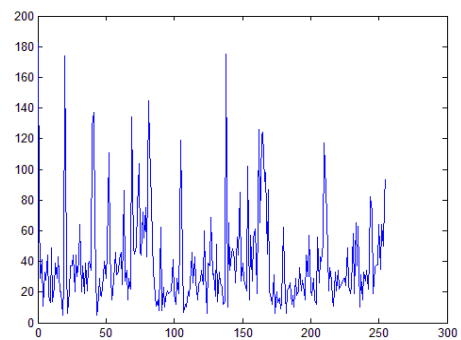
$$F_{10} = \text{مقدار ماکزیمم هیستوگرام} = \max(h) \quad (13)$$

• انرژی هیستوگرام:

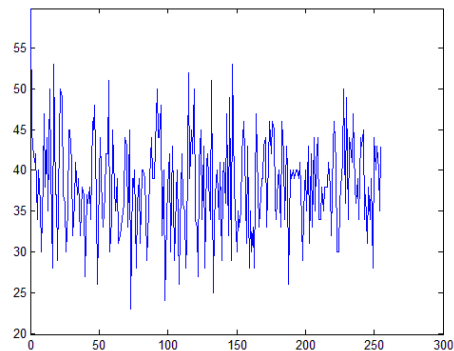
انرژی هیستوگرام ویژگی دیگری است که مورد استفاده قرار می گیرد و به صورت زیر محاسبه می شود:

$$F_{11} = \text{Energy} = \frac{1}{M} \sum_{i=1}^M h(i)^2 \quad (14)$$

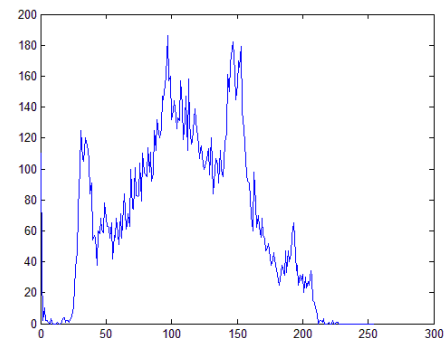
در رابطه (۱۱) مقدار M برابر با 2^m می باشد و m تعداد بیت های هر سمبل می باشد.



شکل (۲). هیستوگرام رشته بیت JPEG ($C_R = 27$)



شکل (۳). هیستوگرام رشته بیت JPEG2000 ($C_R = 27$)



شکل (۴). هیستوگرام رشته بیت تصویر فشرده نشده با فرمت TIFF

2- Entropy

3- feed-forward backpropagation neural network

4- Levenberg-Marquardt

5- neuron

صحت و اعتبار دسته‌بندی‌کننده توسط روابط (۱۷) و (۱۸) محاسبه می‌شود. در این دو رابطه پارامتر C_i برابر با تعداد نمونه‌هایی است که با کلاس i برچسب خورده‌اند و برچسب واقعی آنها نیز i می‌باشد. پارامتر K برابر با تعداد نمونه‌هایی است که به کلاس i تعلق دارند و پارامتر L برابر با تعداد نمونه‌هایی است که با کلاس i برچسب خورده‌اند.

$$A_i = \frac{C_i}{K} \quad (17)$$

$$R_i = \frac{C_i}{L} \quad (18)$$

به‌عنوان مثال صحت و اعتبار کلاس یک به ترتیب، با توجه به روابط (۱۷) و (۱۸) و جدول (۱) عبارتند از:

$$A_1 = \frac{C_1}{C_1+C_{12}+C_{13}} \quad (19)$$

$$R_1 = \frac{C_1}{C_1+C_{21}+C_{31}} \quad (20)$$

صحت و اعتبار کل دسته‌بندی‌کننده به ترتیب طبق روابط (۲۱) و (۲۲) محاسبه می‌شوند:

$$A_{tot} = \frac{C_1+C_2+C_3}{C_1+C_{12}+C_{13}+C_{21}+C_2+C_{23}+C_{31}+C_{32}+C_3} \quad (21)$$

$$R_{tot} = \frac{R_1+R_2+R_3}{3} \quad (22)$$

از شبکه عصبی با الگوریتم پس انتشار خطا برای دسته‌بندی استفاده کرده‌ایم. ولی به‌طور کلی به‌خاطر این که در هر تکرار دو سوم داده‌ها به‌طور تصادفی برای آموزش انتخاب و مابقی برای تست استفاده می‌شود، این الگوریتم در هر بار راه‌اندازی شبکه، مقادیر متفاوت صحت و اعتبار را ارائه می‌دهد. بنابراین سه پیکربندی متفاوت از شبکه را انتخاب و ارزیابی متقابل را سه‌بار تکرار کردیم و از نتایج حاصله میانگین گرفتیم. بدیهی است که این پروسه نه مرتبه فراخوانی الگوریتم روی مجموعه داده‌هایی که دو سوم مقدار اولیه هستند را در بر می‌گیرد.

۷- نتایج شبیه‌سازی

برای آموزش دسته‌بندی‌کننده از پایگاه داده‌ای شامل ۵۱۰ تصویر کدشده توسط JPEG، ۵۱۰ تصویر کدشده توسط JPEG2000 و ۵۱۰ تصویر فشرده‌نشده با فرمت TIFF استفاده کرده‌ایم. تصاویر کدشده دارای نرخ‌های فشرده‌سازی مختلف می‌باشند و از فشرده‌سازی تصاویر خاکستری 8^6 بیت بر پیکسل با ابعاد 512×512 به‌دست آمده‌اند. این تصاویر از پایگاه داده

برای بررسی میزان موفقیت، باید مجموعه یادگیری و مجموعه تست متفاوت از هر کلاس در نظر گرفته شوند. تکنیک ارزیابی متقابل^۱ که اجراکننده آزمایش بر روی حالت‌های جایگردان شده داده‌ها می‌باشد در ادامه شرح داده می‌شود.

۶- ارزیابی متقابل

در نظر گرفتن یک‌سوم از داده‌ها برای تست و باقی‌ماندن دو سوم برای آموزش متداول است. همچنین باید کنترلی صورت داده شود تا نمونه‌های هر یک از کلاس‌ها با تناسب صحیح در مجموعه‌های یادگیری و تست بازنمود شوند. باید توجه داشت که این روش فقط یک اقدام حفاظتی اولیه در برابر ارائه نامساوی نمونه‌ها در مجموعه‌های آموزش و تست را فراهم می‌آورد. روشی کلی‌تر برای کم کردن هر تمایل یک‌طرفه به دلیل نمونه خاص انتخاب شده، تکرار کردن کل پروسه (آموزش و تست) چندین مرتبه با نمونه‌های تصادفی متفاوت است. در هر تکرار دو سوم داده‌ها به‌طور تصادفی برای آموزش انتخاب می‌شود و مابقی برای تست استفاده می‌شود و این پروسه سه‌بار تا انتها تکرار می‌شود و در این حالت هر الگو دقیقاً یک‌بار برای تست استفاده شده است. این پروسه، ارزیابی متقابل سه‌مرتبه‌ای^۲ نامیده می‌شود [۵].

۶-۱- محاسبه صحت و اعتبار جواب

دسته‌بندی‌کننده برای هر کلاس

برای محاسبه صحت^۳ و اعتبار^۴ باید ابتدا ماتریسی $n \times n$ تشکیل داد که n برابر با تعداد کلاس‌ها می‌باشد و به آن ماتریس هم‌رخدادی^۵ گفته می‌شود. این ماتریس در جدول (۱) نشان داده شده است.

جدول (۱). ماتریس هم‌رخدادی 3×3

	کلاس یک	کلاس دو	کلاس سه
کلاس یک (برچسب واقعی)	C_1	C_{12}	C_{13}
کلاس دو (برچسب واقعی)	C_{21}	C_2	C_{23}
کلاس سه (برچسب واقعی)	C_{31}	C_{32}	C_3

- 1- cross validation
- 2- threefold
- 3- Accuracy
- 4- Reliability
- 5- Confusion Matrix

جدول (۴). وزن به دست آمده برای مشخصه‌ها

	W
F ₁	0.1937
F ₂	0.9701
F ₃	0.9939
F ₄	0.9937
F ₅	0.3852
F ₆	0.0190
F ₇	1.0000
F ₈	0.9589
F ₉	1.0000
F ₁₀	0.9054
F ₁₁	0.1983
F ₁₂	0.9954

ماتریس هم‌رخدادی به‌ازای اجرای ۳ بار ارزیابی متقابل سه‌مرتب‌ای در ۰ و با توجه به این ماتریس مقادیر اعتبار و صحت به‌ازای مشخصه‌های انتخاب‌شده در ۰ (۶) نشان داده شده است.

جدول (۵). ماتریس هم‌رخدادی به‌ازای مشخصه‌های انتخاب‌شده

۱۴۸۰	۳۰	۲۰
۱	۱۵۲۸	۱
۳۲	۳	۱۴۹۵

جدول (۶). صحت و اعتبار هر کلاس و مجموع به‌ازای مشخصه‌های انتخاب‌شده

A ₁	R ₁	A ₂	R ₂	A ₃	R ₃	A _{tot}	R _{tot}
٪۹۷	٪۹۸	٪۹۹٫۹	٪۹۸	٪۹۸	٪۹۹	٪۹۸	٪۹۸

۸- نتیجه‌گیری

در این مقاله یک روش پیشنهادی برای محاسبه ماتریس شباهت ارائه نمودیم، که با ترکیبی از معیارهای مجذور فاصله بین دو بردار و زاویه بین دو بردار به‌دست می‌آید. از این روش پیشنهادی برای بهینه کردن الگوریتم انتخاب مشخصه استفاده نمودیم. با توجه به نتایج شبیه‌سازی‌ها مشخص گردید که بدون استفاده از الگوریتم انتخاب مشخصه صحت دسته‌بندی‌کننده ٪۸۹ و اعتبار آن ٪۹۱ است اما با استفاده از الگوریتم انتخاب مشخصه پیشنهادی با وجود این که تعداد مشخصه‌های مورد استفاده برای دسته‌بندی کاهش یافته است، صحت دسته‌بندی‌کننده به ٪۹۸ و اعتبار ٪۹۸ افزایش می‌یابد. در ادامه کار پیشنهاد می‌گردد که ترکیب‌های متفاوتی برای محاسبه ماتریس شباهت بررسی گردد.

SIPI^۱ انتخاب شده‌اند که شامل تصاویر بافت، تصاویر هوایی، تصاویر طبیعی و تصاویر پزشکی می‌باشد و برای کد کردن این تصاویر از نرم‌افزار VCDemo استفاده شده است.

۱۲ ویژگی شرح داده‌شده در بخش ۵ را از رسته‌بیت تصاویر موجود در پایگاه داده محاسبه می‌شوند. برای کاهش حساسیت الگوریتم نسبت به نقطه شروع رسته‌بیت و شیفت آن مقدار هر مشخصه به‌ازای شیفت رسته‌بیت هر نمونه از مقدار یک تا هشت محاسبه شد و مقدار مشخصه اصلی با میانگین‌گیری از این ۸ مقدار به‌دست آمد. در نهایت ۱۵۳۰ بردار مشخصه به‌دست آمد. این بردارهای مشخصه به‌منظور ارتقاء عملکرد دسته‌بندی‌کننده نرمال شدند (رجوع به بخش ۳-۳) و بدون اعمال هیچ‌گونه انتخاب مشخصه به دسته‌بندی‌کننده شبکه عصبی داده شدند. برای دسته‌بندی داده‌ها از شبکه عصبی با یک لایه مخفی استفاده شد. با استفاده از پروسه ارزیابی متقابل سه‌مرتب‌ای، ۳ بار تکرار شد. برای محاسبه مقدار صحت و اعتبار هر مشخصه از نتایج حاصله میانگین گرفتیم. این شبکه با ضریب یادگیری ۰/۰۵ آموزش داده شد. ماتریس هم‌رخدادی به‌ازای اجرای ۳ بار ارزیابی متقابل سه‌مرتب‌ای در ۰ (۲) و با توجه به این ماتریس مقادیر اعتبار و صحت به‌ازای تمام مشخصه‌ها در ۰ و ۰ نشان داده شده است.

جدول (۲). ماتریس هم‌رخدادی به‌ازای تمام مشخصه‌ها

۱۲۲۳	۲۶۵	۴۲
۰	۱۵۳۰	۰
۳۲	۱۶۹	۱۳۲۹

جدول (۳). صحت و اعتبار هر کلاس و مجموع به‌ازای تمام مشخصه‌ها

A ₁	R ₁	A ₂	R ₂	A ₃	R ₃	A _{tot}	R _{tot}
٪۸۰	٪۹۷	٪۱۰۰	٪۷۸	٪۸۷	٪۹۷	٪۸۹	٪۹۱

برای ارزیابی الگوریتم انتخاب مشخصه پیشنهادی، پس از استخراج بردارهای مشخصه و نرمال نمودن آنها، شباهت بین مشخصه‌های استخراج شده، طبق روش گفته شده محاسبه گردید و آنگاه به الگوریتم انتخاب مشخصه شرح داده شده در بخش ۳ اعمال شدند. مشخصه‌هایی که وزن تاثیرگذاری آنها بیشتر از ۰/۵ بدست می‌آید را به‌عنوان مشخصه‌های مطلوب در نظر گرفته و به ماشین دسته‌بندی‌کننده تحویل می‌دهیم. خروجی الگوریتم انتخاب مشخصه به‌ازای مقادیر بهینه $\alpha = 0.2, \beta = 0.8, \delta_1 = 1.2e4, \delta_2 = 3$ در ۰ نشان داده شده است.

٩- مراجع

- [1] P. Suresh, R. M. D. Sundaram, and A. Arumugam, "Feature Extraction in Compressed Domain for Content Based Image Retrieval," International Conference on Advanced Computer Theory and Engineering, Phuket, 2008.
- [2] N. Verma, S. S. Khan, and S. Kant, "Statistical Feature Extraction to Discriminate Various Languages: Plain and Crypt," Scientific Analysis Group, 2003.
- [3] W. H. Vellerling, W. T. Teukolsky, and S. A. Flannery, "Numerical Recipes in C," Second Edition, 1995.
- [4] G. Zhang, "Neural networks for classification: a survey," IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 30, no 4, pp. 451-462, 2000.
- [5] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," Morgan Kaufmann Publishers, 2000.
- [6] H. Liu and Z. Zhao, "Spectral Feature Selection for Supervised and Unsupervised Learning," Proceedings of the 24th International Conference on Machine Learning, 2007.

An Optimized Unsupervised Feature Selection Algorithm

H. R. Kakaei Motlagh, M. Mollazadeh Golmahaleh*, B. Teimorpuor

*Imam Hossein University

(Received: 31/07/2013, Accepted: 12/01/2016)

ABSTRACT

Choosing a feature vector for maximizing the success of a classifier machine is very effective. In this paper, using a combination of different methods to calculate the core function, an unsupervised feature selection algorithm improvement has been proposed. Feature vector obtained by the proposed algorithm, will maximizes output accuracy of backpropagation neural network classifier. In this paper we used case study of standard encoding of images compressed by alternate method and uncompressed images classifying based on their relative bit stream. Standards for classifications are JPEG and JPEG2000 and for uncompressed images is TIFF format. Using this feature vector obtained by the proposed algorithm, classifier accuracy will be about 98%.

Keywords: Feature Vector, Feature Vector Selection, Neural Network, Classification, Image Compressing Standard