

استفاده از تکنیک‌های کاهش بعد تحلیل مولفه‌های اصلی (PCA) و تحلیل جداکننده‌های پارامتری شده (RDA) در رتبه‌بندی نشانگرهای لرزه‌ای

سعیده همت‌پور^۱ و حسین هاشمی^{۲*}

^۱ کارشناس ارشد ژئوفیزیک، دانشکده علوم، دانشگاه آزاد اسلامی واحد تهران شمال، ایران

^۲ استادیار، گروه فیزیک زمین، مؤسسه ژئوفیزیک دانشگاه تهران، ایران

(دریافت: ۸/۱۲/۸۹، پذیرش نهایی: ۱۹/۷/۹۰)

چکیده

هدف از این مقاله انتخاب نشانگرهای بهینه برای آشکارسازی دودکش‌های گازی برای داده‌های بلوک F3 بخش Dutch دریای شمال با استفاده از روش‌های تشخیص الگو براساس معیارهای تحلیل مشخصه تطبیقی (RDA, Regularized Discriminate Analysis) و تحلیل مولفه‌های اصلی (PCA, Principal Component Analysis) است. در علم ژئوفیزیک و مخصوصاً لرزه‌شناسی، که با حجم زیادی داده سروکار دارند و ویژگی‌های مختلفی که می‌توان از آن داده‌ها استخراج کرد، انتخاب مناسب و درست ویژگی‌ها و سپس دسته‌بندی آنها با روش‌های دسته‌بندی، می‌تواند در علم تفسیر داده‌های لرزه‌ای بسیار کار آمد باشد. در رتبه‌بندی نشانگرهای لرزه‌ای همانگونه که در بالا اشاره شد دو روش مورد بررسی در این مطالعه PCA و RDA می‌باشد. با روش PCA نشانگرهای انرژی پنجره‌ای با طول ۸۰ در حول یک نمونه $[-40, 40]$ و انرژی پنجره‌ای با طول ۸۰ در حول یک نمونه $[-40, 120]$ و واکافت موجک طیفی موجک ریکر (فرکانس مرکزی ۶۰ هرتز و پهنای ۲)، به‌عنوان نشانگرهای مطلوب در فضای ویژگی‌ها و نشانگرنوفه تصادفی به عنوان مهمترین نشانگر نامطلوب انتخاب شدند. با توجه به نتایج رتبه‌بندی خواهیم دید که نشانگر تشابه در رتبه بالایی قرار می‌گیرد.

واژه‌های کلیدی: تحلیل مشخصه تطبیقی، تحلیل مولفه‌های اصلی، فضای ویژگی، روش نظارتی، روش غیر نظارتی، ترسیم توزیعی، روش پیشرو، روش پسرو

Using PCA and RDA feature reduction techniques for ranking seismic attributes

Hemmatpour, S.¹ and Hashemi, H.²

¹ MS c. Graduate Geophysics, Faculty of Science, Azad University Tehran Shomal Unit, Iran

² Assistant Professor, Earth Physics Department, Institute of Geophysics, University of Tehran, Iran

(Received: 27 Feb 2011, Accepted: 11 Oct 2011)

Abstract

Optimal attributes are useful in interpretation of seismic data. Two proposed methods are presented in this paper for finding optimal attributes. Regularized Discriminate Analysis (RDA) is based on 2 parameters ϵ , γ which called regularization parameter. The other method is Principal Component Analysis (PCA). In this paper gas chimney detection is defined as the subject of study for ranking relevant attributes. For 4817 samples of both classes i.e., gas chimney and non chimney with 28 attributes which are mentioned in table (1). These attributes have been picked by experienced interpreter. Among all of these attributes some of them such as Similarity (time window: $[-120, -40]$), Similarity (time window: $[40, 120]$), Similarity (time window: $[-40, 40]$), in forward selection algorithm and Similarity (time window: $[-120, -40]$), Similarity (time window: $[-40, 40]$), Energy (time

window: [-120, -40]) in backward selection algorithm in RDA method have the highest ranks. It should be highlighted that because the number of the observations is large so 70% of all observations have been used for train and 30% for test. The discriminate function is:

$$d_k(X_i) = (X_i - \mu_k)^T C_k^{-1}(\lambda, \gamma)(X_i - \mu_k) + Ln |C_k(\lambda, \gamma)| - 2Ln p_k$$

The classification error rate for RDA with $\lambda=0.01$ & $\gamma=0.1$ is 0.09 and for $\lambda=0.1$ & $\gamma=0.1$ is 0.1 and also for $\lambda=0.1$ & $\gamma=0.01$ is 0.09.

In discriminate matrix form which is shown as:

$$d_k(X_i) = (X_i - \mu_k)^T C_k^{-1}(\lambda, \gamma)(X_i - \mu_k) + Ln |C_k(\lambda, \gamma)| - 2Ln p_k$$

C_k is covariance matrix of k-th class, μ_k is mean vector for k-th class and p_k is prior probability of k-th class where $(X_i - \mu_k)^T$ is transpose of $(X_i - \mu_k)$.

In PCA method the principal component obtain by calculating of eigenvectors of covariance matrix and also looking for a transformation with least square error. After these calculations we compare scatter plots of PCA. Selected attributes, PCA method are spectral decomposition with Ricker wavelet (center freq.= 60 (Hz), width=2) and Energy (time window: [-40,40]), Energy (time window: [-40,40]).

For better judgment and selection of optimal attributes we should combine two methods or more and obtain optimal method and also compare different method two by two.

Finally, using pattern recognition method for interpreting of seismic data is suggested.

Key word: Regularized discriminate analysis, Principal Component Analysis, forward selection algorithm, backward selection algorithm, rank, optimal method, covariance matrix.

۱ مقدمه

مقاله مورد بررسی قرار گرفته است عبارتند از RDA و PCA. هاشمی و همکاران (۲۰۰۸) روشی را برای یافتن مرز تصمیم بهینه با ترکیب کلاسه کننده‌های مختلف در مورد مساله دودکش‌های گازی بیان کردند (شکل ۱). هاشمی (۲۰۱۰) در خصوص روش صحیح به کارگیری ابزار تشخیص بافت لرزه‌ای با پارامترهای مناسب در کاربردهای لرزه‌ای مواردی را ارائه نموده که عبارتند از: آگاهی از احتمال پیشین، کمینه اندازه مجموعه آموزش و کاهش نشانگرهای لرزه‌ای زاید.

تکنیکهای دسته‌بندی به دو دسته تقسیم می‌شوند:

۱- روش‌های نظارتی

۲- روش‌های غیر نظارتی

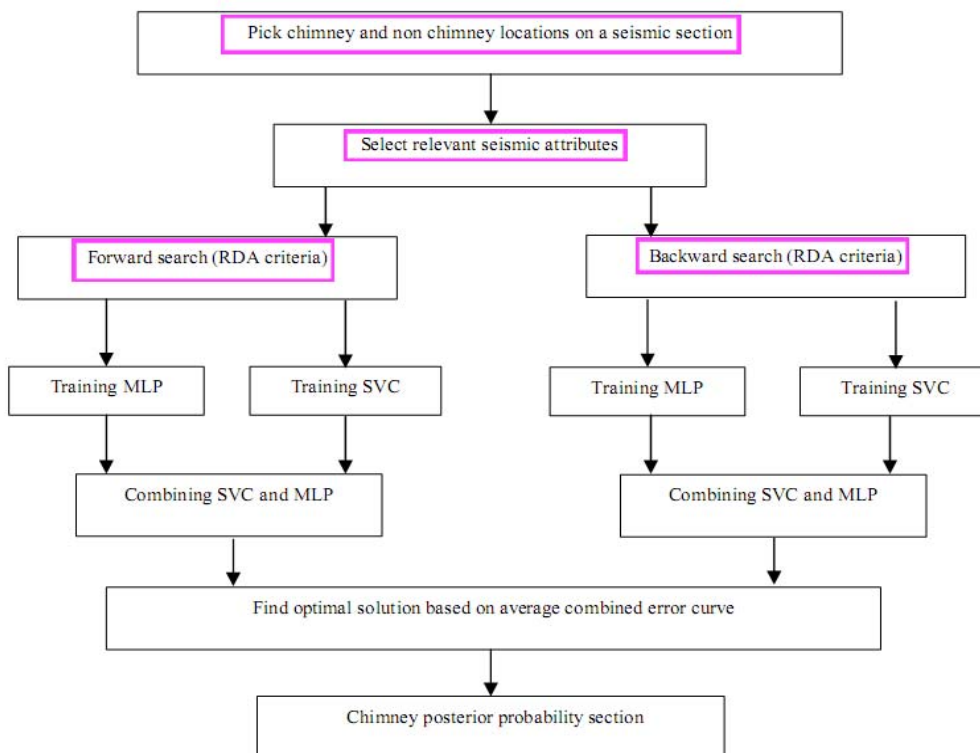
PCA یک روش غیرنظارتی و RDA یک روش

بررسی و انتخاب صحیح نشانگرهای بهینه اغلب سخت و با آزمایش و خطا همراه است. در سال‌های اخیر محققین جهت تعیین نشانگرهای بهینه از روشهایی استفاده کرده‌اند که قبل از شروع طبقه‌بندی نشانگرها را رتبه‌بندی کنند که این کار باعث می‌شود که قبل از طبقه‌بندی نشانگرهای غیر مرتبط حذف شوند و نشانگرهای اضافی وارد سیستم نشوند و فضای ویژگی‌ها دچار ابهام نشود. در سال‌های اخیر به با توجه به پیشرفت تکنولوژی و همچنین حجم زیاد اطلاعات استفاده از روش‌های شبکه‌های عصبی و روشهای جدید تشخیص الگو افزایش یافته است. در مواردی که با حجم زیادی از داده‌ها سر و کار داریم می‌توان با استفاده از روش‌های آماری متعددی به کاهش بعد و دسته‌بندی داده‌ها پرداخت دو روشی که در این

به رتبه‌های به‌دست آمده نشانگرهایی که رتبه‌های بالاتری دارند را انتخاب کرده و به عنوان نشانگرهای بهینه در نظر گرفته می‌شود. در انتها خطای دسته‌بندی دو روش را با هم مقایسه می‌شود.

در این مقاله برای انتخاب نشانگرهای بهینه و جلوگیری از ابهام در کار تفسیر در فضای ویژگی‌ها با بعد بالا از روش پارامتری RDA با پارامترهای متفاوت و همچنین PCA که یکی از ارزشمندترین نتایج جبر خطی می‌باشد و همچنین روش ساده و غیرپارامتری برای استخراج اطلاعات مربوطه از یک مجموعه داده‌های گیج کننده است و تبدیل بهینه در کارهای فشرده‌سازی و کاهش بعد است و خطای میانگین مربعات حاصل از فشرده‌سازی را به حداقل می‌رساند با ترسیم‌های توزیعی مربوط به نشانگرهای مربوطه استفاده شده است.

نظارتی می‌باشد. مشکل اصلی در روش‌های غیرنظارتی انتخاب مناسب و بهینه تعداد گروه‌ها برای دسته‌بندی می‌باشد. استراتژی مورد استفاده در اینجا برای PCA، استفاده از ترسیم توزیعی‌های حاصل از PCA به منظور انتخاب نشانگرهای بهینه می‌باشد بدین صورت که در ترسیم‌های توزیعی نشان داده شده در شکل‌های ۲ و ۳ و ۴ هر کجا که جداشدگی نقاط قرمز (موقعیت دودکش گازی) و آبی (موقعیت غیر دودکش گازی) بیشتر باشد بدین معنی است که این نشانگرها در فضای ویژگی‌ها باقی خواهند ماند، در فضای ویژگی‌ها و جایی که این نقاط هم‌پوشانی زیادی دارند یکی از آنها بایستی از فضای ویژگی‌ها حذف شود و بدین ترتیب کاهش بعد صورت می‌پذیرد. روش دیگر مورد مطالعه روش RDA است که به منظور یافتن رتبه هر نشانگر لرزه‌ای از استراتژی جستجوی پیشرو و پسرو استفاده می‌کند و سپس با توجه



شکل ۱. نمودار موضوعات مورد بحث در این مقاله (هاشمی و همکاران، ۲۰۰۸). قسمت‌های مشخص شده به طور تفصیلی در این مقاله با روش PCA مقایسه شده‌اند.

۲ روش‌های رتبه‌بندی نشانگرها

۱-۲ روش RDA

انتخاب نشانگرهای لرزه‌ای براساس دانش و تجربه مفسر انجام می‌شود. چوپرا و مارفروت (۲۰۰۵) به صورت مبسوط درباره ایده‌های متفاوت به فهرست در آوردن نشانگرها بحث کردند. براساس مطالعات تنر و همکاران (۱۹۹۴) دو دسته‌بندی برای نشانگرهای لرزه‌ای در نظر گرفته شد: نشانگرهای فیزیکی و نشانگرهای هندسی. نشانگرهای فیزیکی اطلاعاتی را درباره فیزیک انتشار موج در زیر سطح (مثل فاز، فرکانس و دامنه) را بیان می‌کند در حالی که نشانگرهای هندسی شکل زیر سطح و هندسه رخدادهای بازتابی را به ما می‌دهد به منظور آشکارسازی اشیاء لرزه‌ای هر دو دسته این نشانگرها مورد نیاز است. به دلیل اینکه ممکن است برخی از نشانگرها دارای مقداری نوفه ناخواسته باشند یا اینکه اطلاعات مفیدی را در رابطه با مسئله مورد نظر در اختیار ما قرار ندهند مفسران در صدد حذف نشانگرهای اضافی و زاید برآمده‌اند و تکنیک‌های رتبه‌بندی فضای ویژگی در کاربردهای تشخیص بافت بروی داده‌های علوم دیگر گسترش یافت. بنابراین ایده تولید زیر مجموعه‌ای از ویژگی‌های اصلی و تخمین معیار جداپذیری کلاس مطرح شد بدین منظور بایستی برخی از موقعیت‌های کلاس‌ها که در اینجا ۲ کلاس (دودکش و غیردودکش) مشخص شود و ویژگی‌هایی برای این نقاط مشخص شود که بیشترین جداشدگی را برای کلاس‌ها به ما بدهد. یک معیار مورد استفاده در این مقاله RDA می‌باشد (فریدمن، ۱۹۸۹). یک مسئله k کلاسی و p نشانگر لرزه‌ای با مقادیر X_{ij} را فرض کنید سپس بردار $X = [X_1, \dots, X_p]$ به عنوان مقادیر لرزه‌ای انتخاب شده در هر نمونه تریس لرزه‌ای تعریف شده است، μ_k بردار میانگین برای K امین کلاس، C_k ماتریس کوواریانس K امین کلاس. P_{μ} احتمال پیشین کلاس k ,

لگاریتم احتمال برای شی x_i برای n برای n پیک برچسب‌دار لرزه‌ای تحت این فرض که توزیع گاوسی است به صورت زیر می‌باشد:

$$d_k(X_i) = (X_i - \mu_k)^T C_k^{-1} (\lambda, \gamma) (X_i - \mu_k) + \ln |C_k(\lambda, \gamma)| - 2 \ln p_k \quad (1)$$

که λ و γ پارامترهای تطبیقی هستند که این پارامترها در عمل بایستی بوسیله تکنیک‌های بهینه‌سازی به دست آیند. ما از دو الگوریتم پیشرو و پسرو برای رتبه‌بندی بر اساس معیار RDA (ون در هیدن و همکاران، ۲۰۰۴) استفاده کرده‌ایم. در روش انتخاب ویژگی پیشرو زیر مجموعه اولیه خالی است و نشانگرها یکی یکی اضافه می‌شوند، اولین نشانگری که انتخاب می‌شود بهترین نشانگر است به همین ترتیب هر چه مقدار خطای کلاسه کننده RDA پایتتر برود نشانگر بدتر می‌شود. در روش پسرو زیر مجموعه اولیه شامل همه نشانگرهاست در هر مرحله یکی از آنها حذف می‌شود و لذا اولین نشانگری که حذف می‌شود بدترین نشانگر و آخرین نشانگری که حذف می‌شود بهترین نشانگر است. لازم به ذکر است که هیچ تضمینی وجود ندارد که نشانگرهای به دست آمده از این دو روش بهترین‌ها هستند بنابراین برای یافتن زیر مجموعه بهینه نشانگرها بایستی زیر مجموعه‌های زیادی تست شوند.

۲-۲ روش PCA

بررسی مولفه‌های اصلی یکی از ارزشمندترین نتایج جبر خطی است زیرا روش ساده و غیر پارامتری استخراج اطلاعات مربوطه از مجموعه‌های گنجاننده است. در PCA ما داده‌ها را از فضایی با بعد بالا به فضایی با بعد پایین می‌بریم در واقع PCA یکی از روش‌های انتخاب ویژگی (Feature selection) است و به منظور کاهش بعد بکار برده می‌شود تا بررسی ویژگی‌ها (Feature) در فضایی با بعد پایین‌تر که ساده‌تر است انجام گیرد. کاری

دسته‌بندی برای کمینه MSE گرفته شوند. در اینجا لازم است به برخی خواص ماتریس کوواریانس پردازیم چون در واقع مساله اصلی ما پیدا کردن مقادیر ویژه ماتریس کوواریانس است.

تعریف کوواریانس در دو بعد به صورت

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

می‌باشد که در واقع ارتباط بین دو بعد که در اینجا X و Y است را بیان می‌کند با محاسبه $\text{COV}(X, X)$ تعریف واریانس حاصل می‌شود. پس واریانس که معیاری از پراکندگی مجموعه داده‌هاست در یک بعد به بررسی این پراکندگی می‌پردازد ولی کوواریانس در دو بعد به بررسی پراکندگی می‌پردازد. حال برای بعدهای بالاتر باز هم با کوواریانس مواجه می‌شویم با این تفاوت که در اینجا دیگر ماتریس کوواریانس را داریم. مثلاً در سه بعد ماتریس کوواریانس به صورت زیر است:

$$C = \begin{pmatrix} \text{cov}(x, y) & \text{cov}(x, y) & \text{cov}(x, z) \\ \text{cov}(y, x) & \text{cov}(y, y) & \text{cov}(y, z) \\ \text{cov}(z, x) & \text{cov}(z, y) & \text{cov}(z, z) \end{pmatrix} \quad (4)$$

که همان‌گونه که مشاهده می‌شود این ماتریس متقارن است و نظر اصلی این ماتریس واریانس داده‌ها را به ما می‌دهد و همچنین این ماتریس یک ماتریس مربعی است. بنابراین می‌توانیم بردارهای مشخصه را برای ماتریس کوواریانس به دست آوریم چون همان‌گونه که از جبر خطی می‌دانیم شرط لازم برای وجود بردارهای مشخصه مربعی بودن ماتریس است.

بردارهای ویژه‌ای ماتریس کوواریانس برهم عمود هستند در اینجا و در مورد PCA بردار ویژه‌هایی مد نظر ما هستند که مطابق با بزرگترین مقادیر ویژه باشند که در واقع مولفه‌های اصلی بردار ویژه‌های با بزرگترین مقادیر ویژه هستند که در امتداد این مولفه‌های اصلی بیشترین

که PCA انجام می‌دهد این است که یک تبدیل خطی Φ به گونه‌ای به دست می‌آورد که بردار ویژگی‌های h -بعدی را به بردارهای ویژگی d -بعدی (که $h < d$) تبدیل می‌کند به طوری که اطلاعات به طور تقریباً کاملی (بیشینه اطلاعات ممکن) نگه داشته می‌شوند و کمینه خطای میانگین مربعات را به دست می‌دهد. تبدیل PCA می‌تواند به وسیله کمینه کردن خطای کمترین مربعات به دست آید. برای دیدن این موضوع فرض کنید بردار ویژگی‌ها $X \in R^d$ (فضای d بعدی) باشد و بعد فضای ویژگی‌های کاهش یافته $Y \in R^h$. اگر بردار ویژگی‌ها $\hat{X} \in R^d$ باشد آنگاه خطای کمترین مربعات می‌تواند به صورت

$$MSE = E \left[\|x - \hat{x}\|^2 \right] \quad (2)$$

بیان شود و در PCA ما به دنبال تبدیلی هستیم که کمترین مقدار خطای مربعات را بدست آوریم یعنی که تبدیل مورد نظر است.

$$T_D = \arg \min_T \left\{ E \left[\|x - \hat{x}\|^2 \right] \right\} \quad (3)$$

اگر از دیدگاه جبر خطی به PCA نگاه کنیم همان‌گونه که گفتیم PCA سعی دارد یک تبدیل خطی T را پیدا کند که کمترین خطای مربعات را به دست دهد و این تبدیل خطی در واقع به گونه‌ای است که $T^T \text{COV}_{x-x} T$ را ماکزیمم کند که در آن COV_{x-x} ماتریس کوواریانس داده‌هایی با میانگین صفر از X است. بنابراین PCA مساله ویژه $\text{COV}_{x-x} V = \lambda V$ را حل کرده و بردار ویژه‌های ستون‌های تبدیل خطی، ماتریس T را تشکیل می‌دهند. داده‌ها در بعد پایین یعنی Y به صورت $Y = (x - \hat{x})^T T$ نشان داده می‌شوند. با توجه به مطالب ذکر شده، MSE حداقل دارای $d-h$ مقدار ویژه است یعنی بردار ویژه بزرگ منطبق با مقدار ویژه بزرگ بایستی برای کاهش بعد و یا به هدف

همبستگی (Correlation) را داریم.

۴ نتایج روش RDA

با توجه به معیار مورد بررسی RDA در روش انتخاب ویژگی پیشرو نشانگرهای تشابه پنجره‌ای [۴۰- و ۱۲۰-] تشابه پنجره‌ای [۴۰ و ۱۲۰] و تشابه پنجره‌ای [۴۰ و ۴۰-] بالاترین رتبه را دارند و در روش انتخاب پسرو تشابه پنجره‌ای [۴۰- و ۱۲۰-] و تشابه پنجره‌ای [۴۰ و ۴۰-] وانرژی پنجره‌ای [۴۰- و ۱۲۰-] بالاترین رتبه را دارند. تعداد نمونه‌ها برای هر دو کلاس ۴۸۱۷ (دودکش گازی و غیر دودکش گازی) و تعداد نشانگرها ۲۸ می‌باشد که در جدول (۱) آمده است. با توجه به اینکه روش RDA روش پارامتری است در این مقاله رتبه‌های حاصله از پارامترهای متفاوت λ و γ نتیجه شده است که در جدول (۲) آورده شده است.

۳ بررسی نتایج حاصله از دو روش RDA و PCA با

استفاده از نرم افزار مطلب

در این بخش به بررسی نتایج به دست آمده روش‌های RDA و PCA با استفاده از نرم افزار مطلب پرداخته شده است. داده‌های مورد استفاده در تشخیص الگو (Pattern Recognition) با استفاده از روش‌های بیان شده برای ارایه بهترین نشانگرها به دو دسته تقسیم شدند که دسته اول با استفاده از ترسیم‌های توزیعی (Scatter plot) به دست آمده از RDA و دسته دوم با استفاده از رتبه‌های به دست آمده از PCA به دست آمده است. پس از به دست آوردن نشانگرهای به دست آمده از دو روش به انتخاب بهترین نشانگرها و حذف نشانگرهای اضافی پرداخته شده است.

جدول ۱. لیست نشانگرهای مورد استفاده در این مطالعه.

1	Reference Time
2	seismic section
3	Energy (time window: [-120,-40])
4	Energy (time window: [-40,40])
5	Energy (time window: [40,120])
6	Polar Dip Variance [-120,-40] step out=1
7	Polar Dip Variance (time window: [-40,40]) step out=1
8	Polar Dip Variance (time window: [40,120]) step out=1
9	Similarity(time window: [-120,-40]),spatial trace position: (-1,2)x(1,-2), FS Min
10	Similarity (time window: [-40,40]), spatial trace position: (-1,2)x(1,-2) FS Min
11	Similarity (time window: [40,120]), spatial trace position: (-1,2)x(1,-2) FS Min
12	Similarity (time window: [-120,-40]), spatial trace position: (1,0)x(0,0) FS Min
13	Similarity (time window: [-40,40]), spatial trace position: (1,0)x(0,0) FS Min
14	Similarity (time window: [40,120]), spatial trace position: (1,0)x(0,0) FS Min
15	Polar Dip
16	Random noise
17	Random noise (time window: [-120,-40])
18	Random noise (time window: [-40,40])
19	Random noise(time window: [40,120])
20	Polar Dip (time shift:-80(ms))
21	Polar Dip (time shift: 80(ms))
22	Curvedness
23	Curvedness (time shift:-80(ms))
24	Curvedness (time shift: 80(ms))
25	Event Asymmetry
26	Mexican Hat Wavelet (25,2)
27	Seismic (low pass:40Hz)
28	Mexican Hat Wavelet(60,2)

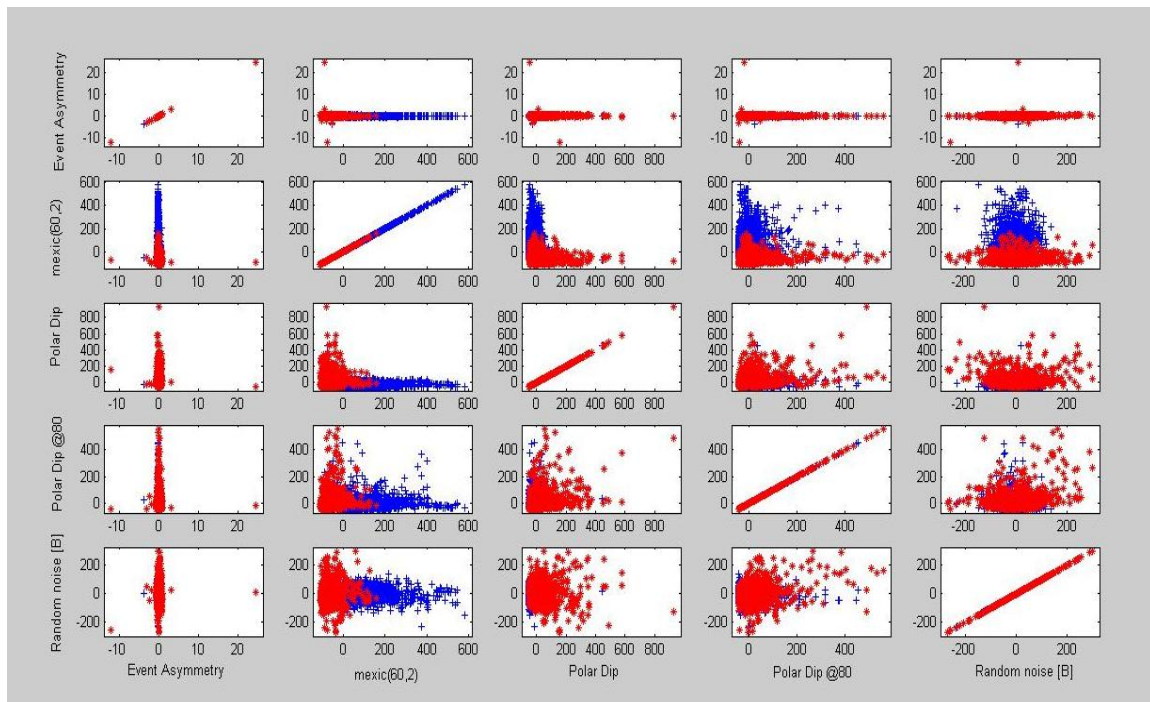
جدول ۲. رتبه‌های به دست آمده از روش RDA با استفاده از روشهای پیشرو و پسرو. در این جدول شماره نشانگرها به جای نام آنها بکار رفته است (رج. جدول ۱).

روش پیشرو			روش پس رو		
$\lambda=0.1, \mu=0.01$	$\lambda=0.1, \mu=0.1$	$\lambda=0.01, \mu=0.001$	$\lambda=0.1, \mu=0.01$	$\lambda=0.1, \mu=0.1$	$\lambda=0.1, \mu=0.001$
۱۳	۱۳	۱۳	۰	۰	۰
۱۱	۱۱	۱۱	۲۷-	۲۷-	۲۷-
۹	۹	۹	۵-	۵-	۵-
۱	۱	۱	۹-	۱۸-	۱۴-
۲۳	۱۰	۲۳	۱۰-	۲۵-	۲۸-
۲۲	۲۲	۲۲	۲۴-	۲۳-	۲۴-
۲۴	۲۴	۲۴	۱۱-	۲۰-	۱۱-
۱۰	۱۰	۱۰	۲۳-	۱۱-	۱۸-
۱۲	۱۲	۱۲	۴-	۶-	۲-
۱۴	۱۴	۱۴	۱۹-	۲۴-	۲۰-
۲۵	۲۵	۲۵	۲-	۱۲-	۱۵-
۲۸	۲۸	۲۸	۱۴-	۱۹-	۹-
۱۵	۱۵	۱۵	۱۳-	۱۷-	۱۹-
۲۱	۱۸	۱۸	۲۱-	۱-	۱۰-
۱۸	۲۱	۱۷	۱۸-	۲۱-	۲۶-
۲۰	۲۰	۲۰	۲۶-	۱۶-	۱-
۱۹	۱۹	۱۹	۱۲-	۷-	۲۱-
۲۶	۱۷	۲۶	۸-	۱۴-	۲۲-
۱۷	۲۶	۲۱	۱-	۲-	۱۶-
۱۶	۱۶	۱۶	۲۵-	۲۸-	۸-
۲	۲	۲	۱۷-	۲۶-	۱۲-
۶	۶	۶	۲۲-	۹-	۷-
۴	۴	۴	۲۰-	۱۳-	۲۳-
۳	۳	۳	۷-	۲۲-	۲۵-
۸	۸	۸	۱۵-	۸-	۱۳-
۷	۷	۷	۱۶-	۱۵-	۶-
۵	۵	۵	۶-	۱۰-	۱۷-
۲۷	۲۷	۲۷	۳-	۳-	۳-

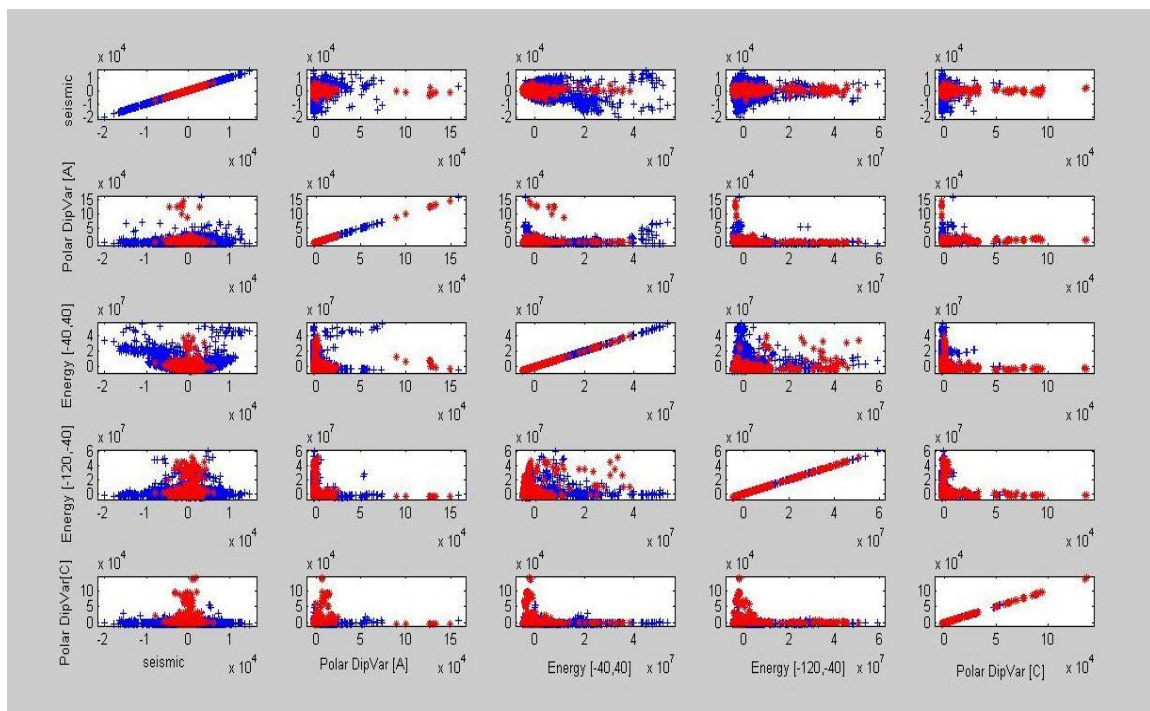
ورودی ترسیم‌های توزیعی PCA برای نشانگرهای مختلف می‌باشد. شکل‌های ۲، ۳ و ۴ نمودارهای نقطه‌ای مربوط به برخی از نشانگرهای این تحقیق را نشان می‌دهد.

۵ نتایج روش PCA

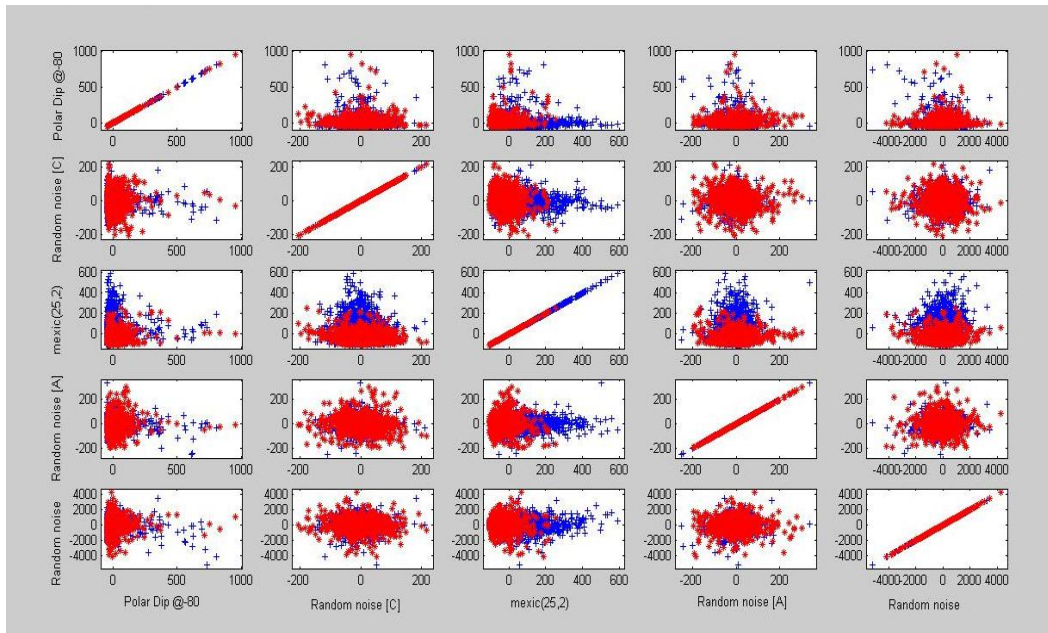
بر خلاف RDA در روش PCA به صورت غیر پارامتری و به صورت گرافیکی به بررسی نشانگرها می‌پردازیم. در واقع وسیله آماری ساده برای کشف ارتباط نشانگرهای



شکل ۲. ترسیم توزیعی حاصل از PCA برای نشانگرهای رخداد نامتقارن، شیب قطبی، نویز تصادفی و واکافت طیفی (موجک ریکر یا کلاه مکزیکیی). جدا شدگی نسبتاً خوب نقاط آبی و قرمز نشان دهنده بهینه بودن نشانگر می باشد که بخوبی توانایی جداسازی محل دودکش را از غیر دودکش دارد.



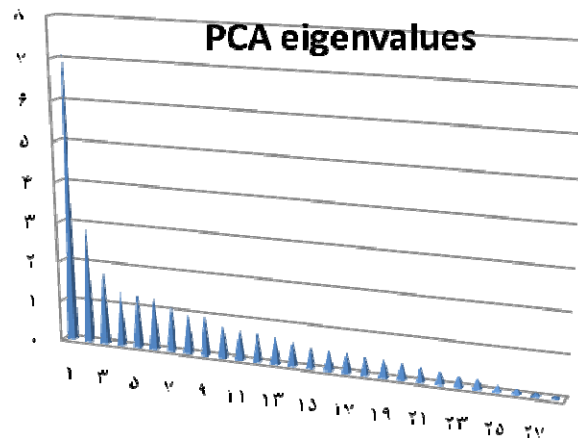
شکل ۳. ترسیم توزیعی حاصل از PCA برای نشانگرهای انرژی، واریانس شیب، شیب قطبی ومقطع لرزه ای. همپوشانی نقاط آبی و قرمز نشان دهنده مرزهای غیر خطی یا زیاد بودن یکی از نشانگرها می باشد که این مساله در این شکل مشهود است.



شکل ۴. ترسیم توزیعی حاصل از PCA برای نشانگرهای شیب قطبی، نویز تصادفی و واکافت طیفی (موجک ریکر یا کلاه مکزیکی). همپوشانی نقاط قرمز و آبی نشان دهنده این است که در نقاط همپوشانی دو نشانگر اطلاعات یکسانی می‌دهند پس وجود هر دو این نشانگرها در فضای ویژگی ضروری نیست.

دو کلاس همپوشانی دارند و نمی‌توان مرز مشخصی برای آن در نظر گرفت از فضای ویژگیها حذف کرد. نمودار مقادیر ویژه به ترتیب نزولی (شکل ۵) با استفاده از نرم افزار مطلب رسم شده است. بیانگر مقادیر ویژه داده های مورد بررسی در این مطالعه می باشد که در تحلیل مولفه های اصلی همانگونه که ذکر شد اولین مولفه اصلی مطابق با بردار ویژه منطبق با بزرگترین مقدار ویژه است.

همانگونه که در اشکال فوق دیده می شود واکافت موجک طیفی موجک ریکر(فرکانس مرکزی ۶۰ هر ترو پهنای ۲) را به عنوان بهترین ویژگی انتخاب شده به دلیل اینکه جداسازی خوبی با بقیه ویژگیها دارد و همچنین انرژی پنجره ای [۴۰ و ۴۰] و انرژی [۴۰ و -۱۲۰] و را به عنوان ویژگیهای مطلوب در فضای ویژگی ها گذاشت. و به طور مثال می توان ویژگی نوفه تصادفی را به دلیل اینکه



شکل ۵. محور عمودی بیانگر مقادیر ویژه و محور افقی محور نشانگرها.

۶ انتخاب روش برتر

با توجه به نتایج به دست آمده و همچنین خطای کلاسه بندی، همانگونه که در جدول ۳ مشاهده می شود خطای حاصل از کلاسه بندی برای روش های RDA و PCA به دست آمده است. اینگونه به نظر می رسد که با توجه به خطای به دست آمده RDA روش بهتری است و هر چه پارامترهای λ, γ کوچکتر باشد خطای دسته بندی کمتر است. از طرفی با توجه به ترسیم های توزیعی حاصل از PCA همانگونه که در شکل ۲ مشاهده می شود نشانگر واکافت موجک طیفی موجک ریکر (فرکانس مرکزی ۶۰ هرتز و پهنای ۲) جداشدگی بسیار خوبی را با بقیه نشانگرها نشان می دهد و بنابراین در فضای ویژگی ها باقی می ماند. ولی در شکل ۴ همانگونه که مشاهده می شود همین نشانگر با نشانگر نوفه تصادفی همپوشانی زیادی دارد که این مساله بیانگر این است که در جاییکه همپوشانی زیاد است دو نشانگر اطلاعات یکسانی می دهند بنابراین اگر یکی از این دو در فضای ویژگیها باقی بماند کافیست. از طرفی همانگونه که در شکل ۲ مشاهده می شود همپوشانی بین نقاط آبی و قرمز وجود دارد ولی در بعضی از قسمت ها جداشدگی خوبی وجود دارد. این مساله به این دلیل به وجود می آید که ممکن است مرزها خطی نباشد و

در دو بعد قادر به مشاهده جداشدگی کلاس ها نباشیم در حالیکه در ۳ بعد بتوان مرزهای ۲ بعدی خوبی مشخص کرد.

۷ نتیجه گیری

نشانگرهای لرزه ای نقش مهمی در آشکارسازی پدیده های لرزه ای ایفا می کنند، بنابراین انتخاب صحیح آنها نقش بسزایی در تفسیر داده های لرزه ای دارد. لذا انتخاب صحیح آنها باعث صرفه جویی در زمان و هزینه می شود.

در این تحقیق به جستجوی نشانگرهای لرزه ای پرداختیم که نقش مهمتری در شناسایی دودکش های گازی دارند نتایج حاکی از این بود که نشانگر تشابه در تعیین مسیر حرکت سیال بسیار مفید است. بنابراین انتخاب نشانگر تشابه به عنوان نشانگر بهینه دور از ذهن نیست. در این مقاله با روش PCA و RDA به جستجوی نشانگرهای موثر بار روی داده های F3 block دریای شمال پرداختیم برای عمومیت بخشیدن به این نتایج در آینده می توان این روش ها را بر روی داده های بیشتری امتحان کرد و نتایج آن را برای انتخاب نشانگرهای مناسب و کافی در شناسایی دودکش گازی به کار برد.

جدول ۳. تعداد نمونه ها در هر کلاس و خطای حاصل از کلاسه بندی با روشهای PCA و RDA.

WITHOUT PREPROCESSING	CLASS 1	CLASS 2	ERROR RATE	CORRECT RATE	PCA	CLASS 1	CLASS 2	ERROR RATE	CORRECT RATE	RDA 0.1&0.1	CLASS 1	CLASS 2	ERROR RATE	CORRECT RATE
CLASS 1	1074	101	0.14	0.86	CLASS 1	1102	73	0.11	0.88	CLASS 1	1060	115	0.1	0.9
CLASS 2	246	1046			CLASS 2	211	1081			CLASS 2	139	1153		

RDA 0.001&0.1	CLASS 1	CLASS 2	ERROR RATE	CORRECT RATE	RDA 0.1&0.01	CLASS 1	CLASS 2	ERROR RATE	CORRECT RATE
CLASS 1	1001	174	0.09	0.91	CLASS 1	1067	108	0.09	0.9
CLASS 2	46	1246			CLASS 2	131	1161		

منابع

- Chopra, S. and Marfurt, K. J., 2005, Seismic attributes: A new historical perspective, *Geophysics* **70**(4), 3 -28.
- Friedman, J. H., 1989, Regularized discriminant analysis, *J. Amer. Statist. Assoc.* **84**, 165-175
- Hashemi, H., 2010, logical considerations in applying pattern recognition techniques on seismic data: precise ruling, realistic solutions, *CSEG Recorder*, **35**, 4.
- Hashemi, H., Duin, R. P. W., Javaherian, A. and de Groot, P., 2008, Gas chimney detection based on improving the performance of combined multilayer perceptron and support vector classifier, *Non linear process in Geophysics*, **15**, 863-871.
- Hashemi, H. and Javaherian, A., 2009, Seismic Attribute Redundancy Reduction Using Statistical Feature Extraction technique, 1st International petroleum conference & Exhibition, EAGE, Shiraz, Iran.
- Lindsay, I. Smith., 2002, A tutorial on principal component analysis, February 26, Department science, University of California, San Diego, December 10.
- Van der Heijden, F., Duin, R. P. W., de Ridder, D. and Tax., D. M. J., 2004, Classification, parameter estimation and estate estimation, 407-413.
- Van der Maaten, L. J. P., 2007, An Introduction to Dimensionality Reduction using Matlab, University of Maastricht.