جداسازی داده های بدون برچسب به کمک برنامه ریزی خطی

حسین موسائی[۱آ] و نازیلا صباغی[ب]
[آ] استادیار، دانشگاه بجنورد، گروه ریاضی، بجنورد، ایران،
[ب] دانشجوی کارشناسی ارشد، دانشگاه بجنورد، گروه ریاضی، بجنورد، ایران

چکیده:

در این مقاله سعی بر این است که داده هایی که بدون برچسب می باشند را به کمک نا مساوی قدرمطلق خوشه بندی کنیم. با استفاده از این روش داده ها ی بدون برچسب به دو دسته تقسیم می شوند. به طوری که بیشتر داده ها درکلاسهای درست و جایگاه مناسب خود قرار گیرند. برای این منظور مدل بدست آمده را به مساله برنامه ریزی خطی تبدیل می‌کنیم و سپس مساله برنامه ریزی خطی را حل می کنیم. آزمایشات عددی نشان دهنده کارایی روش پیشنهادی می باشند.

کلمات کلیدی:

خوشه بندی ; داده های بدون برچسب ; بردار پشتیبان ماشین ; نامساوی قدرمطلق.

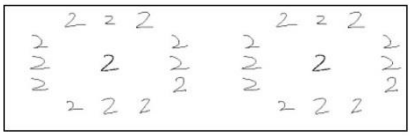
۱- مقدمه:

داده ها و الگوها یکی از شاخص های بسیار مهم در دنیای اطلاعات هستند. در مجموعه ای از داده هایی که طبقه بندی نشده اند، می‌توان ساختار مناسبی یافت که به خوشه بند ی موسوم است .خوشه‌بندی یکی از بهترین روش هایی است که برای کار با داده ها حتی در حالاتی که هیچگونه اطلاعاتی از ساختار داخلی داده ها نداریم و داده ها کاملا بدون برچسب هستند، ارائه شده است . تجزیه و تحلیل خو شه ای رو شی برای گروه‌بندی داده ها با توجه به شباهت یا درجه نزدیکی آنها به یکدیگر ا ست. از این طریق می توان داده ها را به د سته های همگن و متمایز از هم تقسیم کرد. در روش خو شه بندی هیچ کلا سی از قبل وجود ندارد. بلکه ما به دنبال کلاس‌هایی از داده ها هستیم که به هم شباهت دارند و با کمک این شباهت ها می توان رفتارها را بهتر شناسایی کرد و برمبنای آن طوری عمل کرد که نتیجه بهتری حاصل شود [۶].

به بیان دیگر می‌توان گفت که خو شه‌بندی قراردادن داده‌ها در گروه‌هایی ا ست که اع ضای هرگروه از زاویه خا صی شباهت دارند. در نتیجه اعضای یک خوشه به یکدیگر شباهت دارند و با اعضای خوشه های دیگر هیچ شباهتی ندارند.

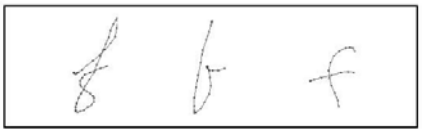
برای مثال شکل های ۱و۲و۳ به خوشه بندی دست خط های متفاوت در افراد اشاره می کند [۲].

[1]Corresponding author: *E-mail address*: hmoosaei@gmail.com



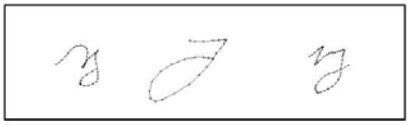
شکل ۱ : دسته بندی مدلهای مختلف

نوشتن عدد 2



شکل ۲ : سه کلاس مختلف برای نوشتن

حرف لاتین f



شکل ۳ : سه کلاس مختلف برای نوشتن

حرف لاتین y

معیار شباهت و هم‌گروهی داده ها در این مقاله فاصله می باشد. یعنی داده‌هایی که به یکدیگر نزدیکترند در یک خوشه قرار می گیرند.

یکی از روش‌هایی که در حال حاضر به صورت گسترده برای جداسازی خوشه ها با محاسبه حداکثر فاصله بین داده های کلاس‌های مختلف مورد استفاده قرار می گیرد، روش ماشین بردار پشتیبان (SVM) است. ایده ی ماشین بردار پشتیبان می کوشد، ابرصفحاتی را پیدا کندکه عمل تمایز نمونه‌های کلاس‌های مختلف داده‌ها را بطور بهینه انجام دهد. بدین گونه که ابرصـــفحه هایی با حداکثر حاشیه را بدست می آوردکه دسته ها را از یکدیگر جدا می‌کند.

در این مقاله می خواهیم با استفاده از نامساوی قدر مطلق در ماشین بردارپشتیبان، داده‌ها ی بدون برچسب را به دو خوشه تقسیم کنیم [۳].

نامساوی قدر مطلق محدب زیر را در نظر بگیرید:

$$\left|x^t w-\gamma\right| \le ۱, \qquad (۱)$$

که در آن $\|.\|$ بیان کننده قدر مطلق است. بردار ستونی x معرف داده در فضای n – بعدی، R^n می باشد. γ فاصله مبدا از صفحه را مشخص می نماید و" ^{t}" ترانهاده است.

نامساوی (۱) را به دو نیم فضای زیر می توان تقسیم کرد:

$$\begin{aligned} x^t w &\le \gamma + ۱, \\ x^t w &\ge \gamma - ۱. \end{aligned} \qquad (۲)$$

هدف این است که بتوانیم به جای استفاده از دو نامساوی (۲) به کمک یک نامساوی داده‌ها را از یکدیگر جدا کنیم. بنابر این به جای نامساوی اول از $x^t w-\gamma \ge ۰$ و به جای نامساوی دوم از $x^t w-\gamma \le ۰$ استفاده می‌کنیم. در واقع هدف ما پیدا کردن ابر صفحه $x^t w=\gamma$ می باشد [۴].

محتویات مقاله شامل موارد زیر است:

در بخش ۲ فرایند خوشه بندی داده های بدون برچسب و تشکیل تابع هدف را بررسی می کنیم. بخش ۳ شامل روش پیشنهادی برای جداسازی داده های بدون برچسب می باشد. در بخش ۴ نتایج عددی را مشاهده می کنیم این نتایج مبین کارایی رویکرد پیشنهادی است. در بخش ۵ نتیجه گیری را بیان می‌کنیم.

شیوه نمادگذاری مقاله حاضر بدین شرح است :

ضرب اسکالر دو بردار ستونی x,y در فضای n-بعدی به صورت $x^t y$ می باشد. بردار ستونی یک با بعد دلخواه را با e و بردارستونی صفر را با 0 نمایش می دهیم. s.t. مخفف" به طوری که "می‌باشد.

۲-خوشه بندی داده های بدون برچسب

فرض کنید داده های بدون برچسب شامل m نقطه در فضای n - بعدی R^n است که با ماتریس $m \times n$، A نشان می دهیم. هر عضو ماتریس A به یکی از کلاس‌های A^+ یا A^- تعلق می گیرد، که در آن A^+ در نیم فضای $x^t w \leq \gamma + ۱$ و A^- در نیم فضای $x^t w \geq \gamma - ۱$ واقع می باشد. لذا داریم:

$$|Aw - e\gamma| \leq e. \qquad (۳)$$

در حقیقت ما به دنبال یافتن حداکثر فاصله دو ابر صفحه $x^t w - \gamma = \pm ۱$ هستیم.

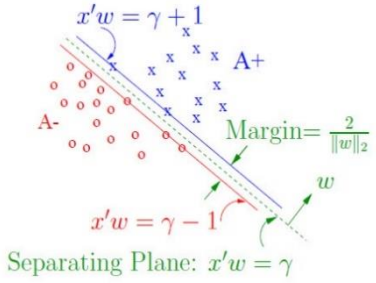
این حداکثر فاصله بین دو صفحه فوق که (حاشیه) نام دارد، یک مزیت برای دستگاه بردار پشتیبان ماشین محسوب می گردد.

همچنین یک رابطه بین w و حاشیه وجوددارد یعنی $\text{حاشیه} = \frac{۲}{\|w\|_۲}$

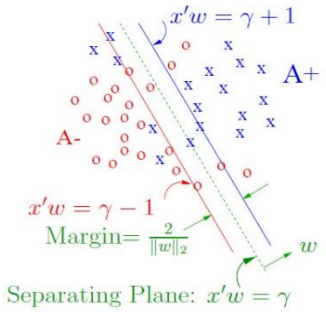
لذا به منظور افزایش حاشیه در تابع هدف باید wرا کاهش دهیم.

از سویی وقتی دوکلاس به شدت خطی و جدا پذیر باشند ،خطای مساله صفر است (شکل ۴). وقتی دو کلاس خطی و جداناپذیر باشند ، مساله دارای خطا می باشد [۵] (شکل ۵) .

بنابراین سعی ما بر آن است که در ضمن کاهش این خطا، حداکثر حاشیه را ایجاد کنیم. همچنین ابر صفحه $x^t w = \gamma$ که وسط دو صفحه فوق می باشد را بیابیم.



شکل ۴ : دو کلاس خطی و جدا پذیر



شکل ۵ : دو کلاس خطی و جدا ناپذیر

به منظور خطی کردن رابطه (۳) می توان این رابطه قدر مطلقی را با کران بالایش محدود کنیم . پس اگر r را کران بالای مورد نظر فرض کنیم، داریم :

$$-r \le Aw - er \le r. \qquad (۴)$$

با توضیحات فوق مساله را به صورت زیر مدل می کنیم :

$$\begin{aligned} \min \quad & -e^t |w| - |\gamma| + ve^t r \\ s.t. \quad & -r \le Aw - e\gamma \le r, \\ & r \le e. \end{aligned} \qquad (۵)$$

که با یک خطی‌سازی پی‌درپی متناهی قابل حل است [۷]. در بخش بعد روشی جدید برای حل مساله (۵) ارائه می‌دهیم.

۳-یادگیری بدون نظارت به کمک برنامه‌ریزی خطی

در این قسمت روشی جدید برای حل مساله (۵) ارائه می دهیم. در روش پیشنهادی مدل را به صورت یک مساله برنامه‌ریزی خطی، تبدیل می‌کنیم و سپس به کمک نرم افزار MATLAB حل می نماییم.

برای این منظور ابتدا توجه داریم که:

$$\begin{aligned} |w| \le k &\rightarrow -k \le w \le k, \\ |\gamma| \le z &\rightarrow -z \le \gamma \le z. \end{aligned} \qquad (۶)$$

پس مساله جدید به صورت زیر مدل می‌شود:

$$\begin{aligned} \min \quad & \cdot\, w + \cdot\, \gamma - e^t k - z + ve^t r \\ s.t. \quad & -r \le Aw - e\gamma \le r, \\ & r \le e, \\ & -k \le w \le k, \\ & -z \le \gamma \le z. \end{aligned} \qquad (۷)$$

لذا مساله زیر را به کمکMATLAB حل می نمائیم و ابر صفحه $x^t w = \gamma$ را تشکیل می دهیم.

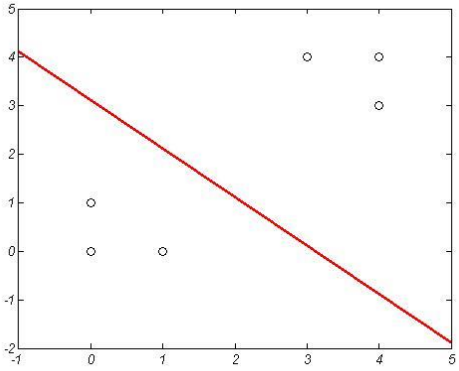
$$
\begin{aligned}
\min \quad & \cdot\, w + \cdot\, \gamma - e^t k - z + \nu e^t r \\
s.t. \quad & Aw - e\gamma \le r, \\
& -Aw + e\gamma \le r, \\
& r \le e, \\
& w \le k, \\
& -w \le k, \\
& \gamma \le z, \\
& -\gamma \le z.
\end{aligned}
\qquad (۸)
$$

در قسمت بعد به بیان نتایج عددی خواهیم پرداخت.

۴-نتایج عددی

در این قسمت روش پیشنهادی را بر روی داده های بدون برچسب اجرا می‌کنیم.
برای اجرای الگوریتم از نرم افزار مطلب ۲۰۱۳ بروی رایانه intel core i5-4210U و حافظه 6GB استفاده شده است.

ابتدا مثالی ۲-بعدی با شش داده را بررسی می کنیم. با اجرای روش پیشنهادی نقاط [0, 0; 0, 1; 1, 0; 4, 4; 3, 4; 4, 3] به دو دسته تقسیم می شود. همانطور که در شکل ۶ مشاهده می شود این جداسازی به طور مناسب انجام شده است.



شکل ۶ : مثال دو بعدی و تقسیم به دو کلاس با روش پیشنهادی

مثال دوم بر روی داده های UCI ، داده مرکز تشخیصی سرطان سینه دانشگاه ویسکانسین آمریکا WDBC و داده Cancer صورت گرفته است [۸و۱].
فرض می‌کنیم داده ها بدون برچسب هستند و آنها را به کمک روش پیشنهادی جدا می‌کنیم و سپس دقت را به دست می‌آوریم.
جدول ۱ نتایج را نشان می دهد.

جدول ۱- نتایج روش پیشنهادی برای داده های WDBC , Cancer

نام داده	تعداد داده ها	تعداد داده های خوشه ها	دقت	زمان(ثانیه)
WDBC	۵۶۹	$A^+ = ۲۵۶$	۵۸,۱۳	۳,۹۰
		$A^- = ۳۱۳$		
Cancer	۳۰۳	$A^+ = ۲۱۶$	۶۰,۰۶	۱,۸۴
		$A^- = ۸۷$		

۵-نتیجه گیری

در این مقاله روشی به کمک نامساویهای قدر مطلق برای جداسازی داده های بدون بر چسب پیشنهاد کردیم. برای حل مدل ایجاد شده ابتدا آن را به صورت برنامه‌ریزی خطی تبدیل کرده و سپس با استفاده از نرم افزار مطلب حل می‌نمائیم. نتایج عددی نشان دهنده کارایی این روش برای جداسازی داده های بدون برچسب می باشد.

مراجع

[1] D.W. Aha and P. M. Murphy, UCI machine learning repository, http://archive.ics.uci.edu/ml, 1992.
[2] K. J. Anil, " Data clustering: 50 years beyond K-means" *Pattern recognition letters* Vol. 31, No.8, pp 651-666, 2010.
[3] G. M. Fung and O.L. Mangasarian, "Data selection for support vector machine classifiers" *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. USA, pp. 64-70, 2000.
[4] G. M. Fung and O.L. Mangasarian, "Semi-superyised support vector machines for unlabeled data classification" *Optimization methods and software*, Vol. 15, No. 1, pp. 29-44, 2001.
[5] G. M. Fung and O.L. Mangasarian, "Unsupervised and Semisupervised Classification via Absolute Value Inequalities", Journal of Optimization Theory and Applications, Vol. 168. No. 2, pp. 551-558, 2016.
[6] C. Ma and G. Gan and J. Wu, *Data clustering: theory, algorithms, and applications*. SIAM, Vol. 20, 2007.
[7] O. L. Mangasarian, "Unsupervised classification via convex absolute value inequalities" *Optimization*, Vol. 64, No. 1, pp. 81-86, 2015.
[8] W. H. Wolberg and W. N. Street and O. L. mangasarian, WDBC: Wisconsin Diagnostic Breast Cancer Database, Computer Sciences Department University of Wisconsin Madison, ftp://ftp.cs.wise.edu/math-prog/cop-dataset/machine-learn/cancer/WDBC,1995.