13th سیزدهمین کنفرانس بین المللی فناوری اطلاعات و دانش
۲۹ آذر ماه الی ۱ دی ماه ۱۴۰۱
13th International Conference on Information & Knowledge Technology
"ایران هوشمند در پرتو فناوری اطلاعات و دانش"

# بررسی مقاله ماشین بردار پشتیبان تودرتو

مهانیان مسعود

**چکیده – ماشین های بردار پشتیبان (*SVMS*) در ابتدا برای طبقه بندی های باینری طراحی شدند. در مورد چند طبقه‌بندی، معمولاً به دوتایی تبدیل می‌شوند، جایی که معمولاً مناطق غیرقابل طبقه‌بندی وجود دارد. برای غلبه بر این اشکال، روش جدیدی به نام ماشین بردار پشتیبان تودرتو[1] (*NSVMS*) برای طبقه‌بندی چندگانه در این مقاله بررسی شده. و معایب و مزایای آن بیان شده سات. ایده به شرح زیر است: ابتدا، ابرصفحه های بهینه را بر اساس الگوریتم یک در مقابل یک بسازید. ثانیاً، اگر نقاط داده‌ای در ناحیه میانی غیرقابل طبقه‌بندی وجود دارد، آنها را برای ساخت ابرصفحه‌های بهینه با همان فراپارامترها انتخاب کنید. ثالثاً، مرحله دوم را تکرار کنید تا زمانی که هیچ نقطه داده ای در مناطق غیرقابل طبقه بندی وجود نداشته باشد یا مناطق ناپدید شوند**[1].

کلید واژه- ماشین های بردار پشتیبانی; دستگاه بردار پشتیبانی حداقل مربعات; الگوریتم یک در مقابل یک؛ FLS-SVM; ماشین بردار پشتیبانی تودرتو. چند طبقه بندی

## 1- مقدمه

ماشین‌های بردار پشتیبان استاندارد (SVM) [1] در ابتدا برای طبقه‌بندی‌های باینری طراحی شدند. متأسفانه، بسیاری از برنامه های کاربردی شامل مسائل چند طبقه بندی هستند که معمولاً به موارد باینری تبدیل می شوند. تاکنون روش‌های مختلفی برای تجزیه و بازسازی مسائل چند طبقه‌بندی پیشنهاد شده است که می‌توان به دید خوبی از این روش‌ها نوشته ریفکین اشاره کرد.

مزایای ماشین های بردار پشتیبان عبارتند از:

- موثر در فضاهای با ابعاد بالا
- در مواردی که تعداد ابعاد بیشتر از تعداد نمونه ها باشد همچنان موثر است.
- از زیرمجموعه ای از نقاط آموزشی در تابع تصمیم استفاده می کند. (به نام بردارهای پشتیبان)
- چند منظوره: توابع کرنل مختلفی را می توان برای تابع تصمیم مشخص کرد. هسته های مشترک ارائه شده است، اما امکان تعیین هسته های سفارشی نیز وجود دارد.

معایب ماشین های بردار پشتیبان عبارتند از:

- اگر تعداد ویژگی‌ها بسیار بیشتر از تعداد نمونه‌ها است، در انتخاب توابع کرنل از تناسب بیش از حد خودداری کنید و اصطلاح منظم‌سازی[2] بسیار مهم است.

[1] NESTING SUPPORT VECTOR MACHINTE
[2] regularization

- SVM ها به طور مستقیم تخمین های احتمال را ارائه نمی دهند، این ها با استفاده از اعتبارسنجی متقابل پنج برابری هزینه محاسبات بیشتر می‌شوند.

در این مقاله سعی شده که از الگوریتم های یک در برابر یک که یک الگوریتم باینری می‌باشد استفاده شود. در مورد الگوریتم یک در برابر یک SVC و NuSVC، طرح بندی ویژگی ها کمی بیشتر درگیر است. در مجموع، n_classes * (n_classes - 1) / 2 طبقه‌بندی‌کننده ساخته می‌شوند و هر کدام داده‌ها را از دو کلاس آموزش می‌دهند.

## 2- تئوری مقاله

### 2-1- ماشین بردار پشتیبانی استاندارد

فرمول اصلی Vapnik، طبقه بندی کننده SVM باینری به شرح زیر است:

$$D(x) = w^T \psi(x) + b$$

برای جداسازی خطی داده ها در فضای ویژگی، تابع تصمیم شرایط زیر را برآورده می کند:

$$y_i (w^T \psi(x_i) + b) \geq 1 \qquad for \quad i = 1, \cdots, l$$

برای تعیین ابرصفحه جداکننده بهینه که دارای حداکثر حاشیه بین دو کلاس است، می‌توانیم مسئله بهینه‌سازی زیر را فرموله کنیم:

$$y_i (w^T \psi(x_i) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

برای بدست آوردن صفحه جداسازی بهینه، باید مسئله بهینه سازی زیر را حل کنیم

$$\min_{w,b,\xi} J(w,b,\xi) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^{l} \xi_i ,$$

s. t.

$$y_i (w^T \psi(x_i) + b) \geq 1 - \xi_i \qquad for \quad i = 1, \cdots, l$$
$$\xi_i \geq 0 \qquad for \quad i = 1, \cdots, l$$

که γ پارامتری است که مبادله بین حداکثر حاشیه و حداقل خطای طبقه بندی را تعیین می کند.

لاگرانژی مربوطه عبارت است از:

$$L(w,b,\xi;\alpha,\beta) = J(w,b,\xi) - \sum_{i=1}^{l} \alpha_i \{ y_i [w^T \psi(x_i) + b] - 1 + \xi_i \}$$
$$- \sum_{i=1}^{l} \beta_i \xi_i$$

که با مشتق گیری از لاگرانژ و حل آن به پارامترهای مورد نظر میرسیم:

$$Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \psi(x_i)^T \psi(x_j)$$

$$\sum_{i=11}^{l} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq \gamma$$

### 2-2- تابع هسته

تابع هسته می تواند یکی از موارد زیر باشد:

- خطی
- چند جمله ای
- rbf
- سیگموئید

### 2-3- One-against-One Algorithm

در این الگریتم در مجموع به n_classes * (n_classes - 1) / 2 کلاس طبقه‌بندی‌کننده ساخته می‌شوند و هر کدام داده‌ها را از دو کلاس آموزش می‌دهند و هدف یافتن داده های کلاس بندی نشده است.

### 2-4- شرح الگوریتم

در مرحله اول، ابرصفحه های را n_classes * (n_classes - 1) / 2در فضای ویژگی بر اساس رویکرد یک در برابر یک بسازید.

ثانیا، نقاط داده را در منطقه بیابید.

ثالثاً، از نقاط داده در ناحیه غیرقابل طبقه‌بندی به تنهایی برای ساختن هایپرصفحه با همان فراپارامترها استفاده کنید.

در نهایت مرحله دوم و سوم را تکرار کنید تا زمانی که وجود داشته باشدهیچ نقطه داده ای در ناحیه میانی یا منطقه ناپدید نمی شود.

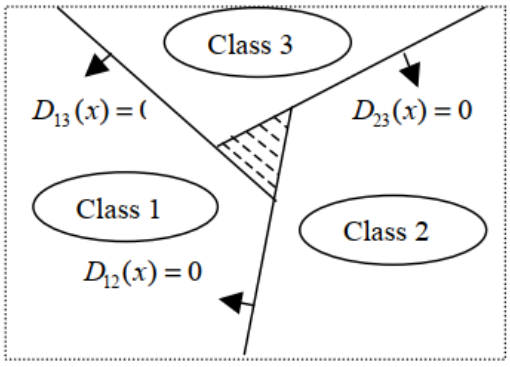
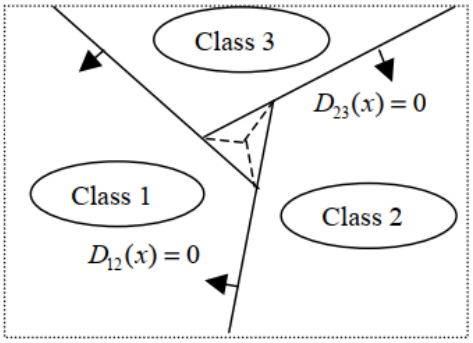



Figure 1 One-against-One



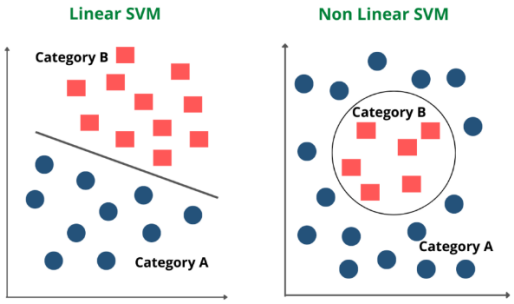

## 3- نتیجه‌گیری

### 3-1- مزایا این روش:

- کارکرد فوق العاده در حل مسائل پیچیده
- سادگی پیاده سازی
- مدیریت داده‌های زیاد

### 3-2- معایب این روش:

- کرنل به صورت خطی می‌باشد و نمیتواند داده‌هایی با شکل زیر را حل کند.




- در مسائل پیچیده به سادگی Overfitt می‌شود.
- نسبت به نویز و داده خطا بسیار ناپایدار می‌باشد.
- هایپر پلن ها حتما باید در تداخل باشند و یک فظای هندسی بسازند.
- روش های بهتری برای بررسی عدم کلاس بندی درست موجود است.
- مشکلات نامتعادل بودن و ایجاد وزن در داده‌ها بررسی شده است.

### 3-3- نتایج پیاده سازی

پیاده سازی ها در دو پایگاه داده انجام شده است و در نهایت با روش‌های دیگر یادگیری ماشین (یادگیری رگرسیون لاجستیک، KNN، بیز ساده لوحانه و درخت تصمیم) مقایسه شده است که به شرح زیر است:

توجه داشته باشید که در کد بهترین KNN محاسبه شده است.

پایگاه داده تیروئید:

دارای 1800 داده می‌باشد.

ابتدا در داده آموزش 32 داده اشتباه کلاس بندی شده است که در مرحله بعدی کلاس بندی این 32 کلاس تصحیح شده است.

مرحله اول:

| | واقعی | | | |
|---|---|---|---|---|
| | کلاس 1 | کلاس 2 | کلاس 3 | کلاس 4 |
| کلاس1 | 1876 | 0 | 0 | 0 |
| کلاس 2 | 4 | 12 | 0 | 0 |
| کلاس3 | 10 | 2 | 0 | 0 |
| کلاس4 | 23 | 0 | 0 | 0 |

مرحله دوم(که تمامی داده ها اشتباه کلاس بندی شده اصلاح می‌شود.):

| | واقعی | | | |
|---|---|---|---|---|
| | کلاس 1 | کلاس 2 | کلاس 3 | کلاس 4 |
| کلاس1 | 2 | 0 | 0 | 0 |
| کلاس 2 | 0 | 3 | 1 | 0 |
| کلاس3 | 0 | 2 | 7 | 0 |
| کلاس4 | 0 | 0 | 0 | 19 |

که این نشان از دقت بالای الگوریتم می‌دهد.

| | الگوریتم مورد نظر | یادگیری رگرسیون لاجستیک | مناسب KNN | بیز ساده لوحانه | درخت تصمیم |
|---|---|---|---|---|---|
| دقت | 0.99 | 0.99 | 0.99 | 0.93 | 0.99 |
| میانگین کلان | 0.99 | 0.71 | 0.79 | 0.56 | 0.81 |
| میانگین وزنی | 0.99 | 0.98 | 0.98 | 0.95 | 0.99 |

f1-score

پایگاه داده قیمت موبایل:

دارای 2000 داده می‌باشد.

با توجه به این که در این پایگاه داده تعداد کلاس ها تقریبا برابر است پس بررسی آن مورد توجه قرار گرفته است:

ابتدا در داده آموزش 41 داده اشتباه کلاس بندی شده است که در مرحله بعدی کلاس بندی این 41 کلاس تصحیح شده است.

با توجه به پراکندگی متقارن بررسی دقت روش مناسبی است.

| | الگوریتم مورد نظر | یادگیری رگرسیون لاجستیک | مناسب KNN | بیز ساده لوحانه | درخت تصمیم |
|---|---|---|---|---|---|
| دقت | 0.99 | 0.96 | 0.62 | 0.82 | 0.87 |

## مراجع

[1]Liu, Bo, Zhifeng Hao, and Xiaowei Yang. "Nesting algorithm for multi-classification problems." Soft Computing 11 (2007): 383-389.

[2]S. Abe and T. Inoue, “Fuzzy support vector machines for multiclass problem”, In Proceedings of 10th European Symposium on Artificial Neural Networks (ESANN”2002), Bruges, Belgium pp. 113-118, April 2002.