دانشکده علوم ریاضی
گروه ریاضی کاربردی

پایان نامه

برای دریافت درجه کارشناسی ارشد در رشته
ریاضی کاربردی، گرایش تحقیق در عملیات

عنوان

طبقه‌بندی داده‌ها با استفاده از برنامه‌ریزی ریاضی

استاد راهنما

جناب آقای دکتر تقی زاده

استاد مشاور

جناب آقای دکتر قنبری

پژوهشگر

آلاله مسکوکی

بهمن ماه ۱۳۹۰


بسم الله الرحمن الرحيم

بسمه تعالی .

مشخصات پایان نامه تحصیلی دانشجویان .

دانشگاه فردوسی مشهد

عنوان پایان نامه: طبقه‌بندی داده‌ها با استفاده از برنامه‌ریزی ریاضی

نام نویسنده: آلاله مسکوکی

نام استاد(ان) راهنما: آقای دکتر تقی‌زاده کاخکی

نام استاد(ان) مشاور: آقای دکتر قنبری

دانشکده : علوم ریاضی	**گروه: ریاضی**	**رشته تحصیلی: ریاضی کاربردی-تحقیق در عملیات**
تاریخ تصویب: ۱۳۹۰/۳/۱۶		**تاریخ دفاع: ۱۳۹۰/۱۱/۵**
مقطع تحصیلی: کارشناسی ارشد ●	**دکتری ○**	**تعداد صفحات: ۹۷**

چکیده پایان نامه :

پیشرفت‌های اخیر در جمع‌آوری داده‌های دیجیتال و تکنولوژی‌های ذخیره سازی باعث پدید آمدن توده‌های حجیم داده و گسترش آن شده است. پدید آمدن سریع و رو به گسترش حجم عظیم داده‌ها در زندگی امروزی کاربردهای وسیعی را در زمینه‌های مختلف علمی، مهندسی و پزشکی به وجود آورده است. استخراج رابطه پنهان میان این انبوه داده و خلاصه کردن آن‌ها به صورت ابتکاری که قابل فهم و در نتیجه قابل استفاده برای صاحبان داده باشد، یکی از مهم‌ترین اهداف طبقه‌بندی به شمار می‌رود.

در این پایان‌نامه، به معرفی مسأله طبقه‌بندی می‌پردازیم و به روش‌های مختلف برای حل این مسأله، به‌خصوص روش‌های برنامه‌ریزی ریاضی، اشاره می‌کنیم. در ادامه، برخی مدل‌های برنامه‌ریزی خطی، اعداد صحیح و غیرخطی ارائه شده در مقالات مختلف را بررسی می‌کنیم. در پایان، یک مدل خاص برنامه‌ریزی اعداد صحیح موجود، بررسی و پیاده‌سازی شده و برای بهبود آن پیشنهاداتی ارائه می‌دهیم. نتایج محاسباتی، کارائی اصلاحات پیشنهادی را تأئید می‌کند.

کلید واژه:

۱. طبقه‌بندی داده‌ها

۲. یادگیری نظارتی

۳. برنامه‌ریزی ریاضی

امضای استاد راهنما:

تاریخ: ۱۳۹۰/۱۱/۲۹

فهرست

صفحه

فصل ۱

فصل ۲

فصل ۳

فصل ۴

فصل ۵

پیش‌گفتار

طبقه‌بندی داده‌ها یکی از مسایل اساسی در داده‌کاوی است. روش‌های طبقه‌بندی به عنوان وسیله‌ای برای تصمیم‌گیری در بسیاری از زمینه‌ها از جمله پزشکی، شناسایی الگو، اقتصادی و بانکداری، کنترل کیفیت و غیره استفاده می‌شود. در این مسأله با تعریف یک تابع به عنوان تابع جداکننده، به تعیین عضویت داده‌های ناشناخته در گروه‌های مشخص بر پایه ویژگی‌ها و اطلاعاتی که از داده‌ها در دست است، می‌پردازیم. برای این منظور ابتدا تعدادی داده که عضویت آن‌ها درگروه ها مشخص است برای تعیین وزن‌های تابع جداکننده با هدف مینیمم کردن داده‌های نادرست طبقه‌بندی شده به کار می‌رود که به آن مرحله آموزش گفته می‌شود. سپس در مرحله آزمون، داده‌های معلوم جدیدی (که در آموزش استفاده نشده‌اند) را برای تعیین عضویت درگروه‌های مفروض در تابع قرار داده و دقت تابع اندازه‌گیری می‌شود.

فصل اول به معرفی شناسایی الگوها در قالب یک مثال ساده می‌پردازد. در این فصل با تعداد، پیچیدگی، گوناگونی و بزرگی برخی زیرمسأله‌های شناسایی الگو آشنا می‌شویم. این زیرمسأله‌ها به ندرت به تنهایی مورد توجه بوده و همواره در ارتباط تنگاتنگ با مسائل دیگر هستند. در این میان، جایگاه مسأله طبقه‌بندی داده را در فرایند شناسایی الگوها مشخص می‌کنیم.

در فصل دوم به معرفی مسأله طبقه‌بندی (یادگیری نظارتی) می‌پردازیم و به برخی از کاربردهای مهم و مطالعات انجام شده در این زمینه اشاره می‌کنیم.

فصل سوم را با بیان نظریه تصمیم بیز شروع می‌کنیم. این روش برای حالت ایده‌آلی است که ساختار تابع احتمال وقوع برای گروه‌ها کاملاً شناخته شده است. اگرچه که چنین وضعیتی به ندرت در عمل اتفاق می‌افتد، اما این روش به ما این امکان را می‌دهد که جداکننده بهینه را به دست آورده و جداکننده های دیگر را با آن مقایسه کنیم. در ادامه، به معرفی روش شبکه‌های عصبی به طور مختصر می‌پردازیم. برخی از ایده‌های جداکننده های خطی را می‌توان به یک کلاس از الگوریتم‌های قدرتمند برای آموزش شبکه‌های عصبی تعمیم داد. برخی از روش‌های طبقه‌بندی، همانند الگوریتم‌های مبنی بر درخت‌ها، توسط قوانین منطقی بیان می‌شود. در اینجا، نحوه عملکرد یک الگوریتم درختی را در قالب یک مثال ساده بیان می کنیم. بررسی روش‌ها را با معرفی «ماشین بردار پشتیبان» که یک جداکننده ابرصفحه ای به صورت یک مدل برنامه‌ریزی ریاضی است پایان می‌دهیم.

فصل چهارم به طور خاص، به روش برنامه‌ریزی ریاضی اختصاص یافته است. به طور کلی، مسائل طبقه‌بندی را بر پایه تعداد کلاس‌ها به دو دسته «دو گروهی» و «چندگروهی» تقسیم می‌کنند. روش‌های طبقه‌بندی دوگروهی را می‌توان با استفاده از الگوریتم‌هایی برای مسائل چندگروهی نیز به کار برد. اساس مدل های برنامه‌ریزی خطی، تولید ابرصفحه های جداکننده در فضای ویژگی‌هاست. مدل های بسیاری برای حل مسأله دوگروهی با این روش تاکنون مورد مطالعه قرار گرفته‌اند. این در حالی است که تعداد مدل‌های چندگروهیِ ارائه شده زیاد نیست. در این فصل به بررسی برخی مدل‌های خطی، عددصحیح و غیرخطی رایج تر قدیمی و همچنین مدل‌های جدیدتر برگرفته شده از آن ها به همراه خلاصه‌ای از مقایسات انجام شده برای هر دو دسته از مسائل می‌پردازیم.

اینکه «هیچ روشی وجود ندارد که بهترین عملکرد را نسبت به دیگر روش‌ها بر روی تمام انواع الگوها داشته باشد» یک اصل پذیرفته شده است. اگر جداکننده‌ای بر روی یک مثال، بهتر از دیگری عمل می‌کند، می‌تواند به دلیل تناسب ساختار جداکننده با نوع خاصی از الگوها باشد. از طرفی، اکثر روش های مطرح شده برای طبقه بندی چندگروهی با مثال‌های کوچک و متوسط ارزیابی شده است. عملکرد چند روش طبقه بندی بر روی الگوهایی از ابعاد بزرگ با خصوصیات مختلف آماری با یکدیگر مقایسه شده است. نتایج این بررسی، دقت بالای روش برنامه‌ریزی ریاضی را برای حل مسأله طبقه‌بندی نسبت به روش‌های دیگر نشان می‌دهد که در پایان فصل چهارم به آن اشاره شده است.

ابرمکعب‌ها، نوع ساده‌ای از جداکننده های چندوجهی هستند. در فصل پنجم، یک مدل برنامه‌ریزی اعداد صحیح خاص برای طبقه‌بندی به همراه یک الگوریتم تکراری بررسی شده است. از نقاط قوت این روش، دقت بالای آن می‌باشد. همان‌طور که می‌دانیم، مشکل اساسی مدل های برنامه‌ریزی با اعداد صحیح زمان زیاد محاسباتی برای داده‌های با ابعاد بزرگ است که کارآیی این روش‌ها را کاهش می‌دهد. در ادامه این فصل، نوع تغییریافته‌ای از مدل اصلی را پیشنهاد می‌کنیم که به طور قابل‌ملاحظه‌ای زمان آموزش را کاهش می‌دهد.

بسمه تعالی

Graduate Studies Thesis\Dissertation Information
Ferdowsi University of Mashhad

Title of Thesis\Dissertation: Data Classification Using Mathematical Programming

Author: Alaleh Maskooki
Supervisor(s): Dr. Taghizadeh Kakhki
Advisor(s): Dr. Ghanbari

Faculty: Mathematical Sciences	**Department:** Mathematics	**Specialization:** Applied Mathematics-Operations Research

Approval Date: 16/3/1390	**Defense Date**: 5/11/1390
M.Sc. ● **Ph.D.** ○	**Number of Pages:** 97

Abstract:

Recent advances in digital data collection and storage technologies brought about massive data sets. The emergence and exponential growth of data avalanche nowadays, created wide applications in various fields of science, engineering and medicine. Extracting hidden relationship between data, and summarizing them innovatively, to be understandable and thus practical for data owners, is one of the most important tasks in data classification.

In this thesis, we introduce the data classification problem and refer to different methods, particularly mathematical programming methods, for solving this problem. Some linear, integer and nonlinear programming models, existing in literature, are studied. A specific mixed integer programming based model is implemented and discussed. We provide some recommendation to improve the efficiency of the method. Computational results confirm the practicability of the recommended algorithm.

Signature of Supervisor:	**Key Words:**
Date: 18 Feb. 2012	1. Data classification problem 2. Supervised learning 3. Discriminant analysis 4. Mathematical programming

Ferdowsi University of Mashhad
Department of Applied Mathematics

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in applied mathematics

Title

Data Classification Using Mathematical Programming

Supervisor:
Dr. Taghizadeh

Advisor:
Dr. Ghanbari

Author:
Alaleh Maskooki

January 2011

فهرست

صفحه

فصل ۱

فصل ۲

فصل ۳

فصل ۴

فصل ۵

فصل ۱

معرفی شناسایی الگو[1]

۱.۱ مقدمه

ما به سادگی توانایی شناسایی یک چهره یا صدا را داریم، حروف الفبای دستنویس را می‌خوانیم، با لمس کردن اشیاء در جیب خود آن‌ها را از یکدیگر متمایز می‌سازیم و یا نوع ماده‌ای را با بوییدن تشخیص می‌دهیم. همه این‌ها مربوط به فرآیندهای پیچیده شناسایی الگوهاست، که به طور شگفت انگیزی با دقت بالا در ذهن انسان انجام می‌گیرد. این امری طبیعی است که ما به دنبال طراحی و ساخت ماشین‌هایی باشیم که قادر به شناسایی و طبقه‌بندی الگوها باشند.

در آغاز، به معرفی علم شناسایی الگو می‌پردازیم. برخی زیر مسأله‌ها و اصطلاحات رایج را در قالب یک مثال بیان کرده و جایگاه طبقه‌بندی داده را در این علم مشخص می‌کنیم. این مثال از [۲۸] برگرفته شده‌است.

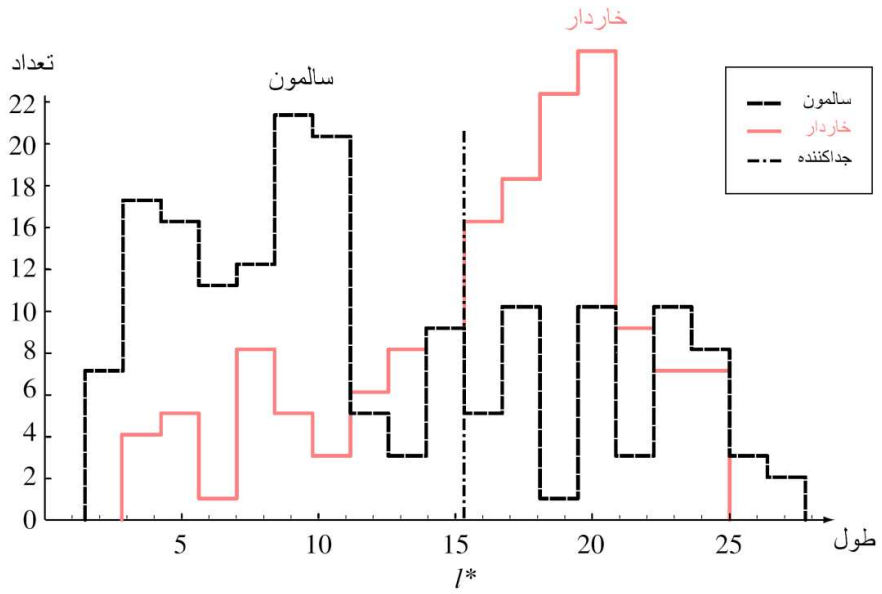
۲.۱ یک مثال کاربردی

فرض کنید یک کارخانه بسته‌بندی ماهی تصمیم دارد فرآیند دسته‌بندی ماهی‌های دریافتی را که بر روی نوار نقاله قرار می‌گیرند، بر اساس گونه آن‌ها به صورت ماشینی انجام دهد. به عنوان یک پروژه آزمایشی، با استفاده از یک حسگر نوری، قرار است ماهی‌های سالمون از ماهی‌های خاردار جدا شوند. به این منظور، توسط یک دوربین تعدادی عکس از هر دو گونه گرفته می‌شود و برخی از تفاوت‌های فیزیکی این دو گونه ماهی مثل طول، عرض، وزن، شکل، تعداد باله‌ها و موقعیت دهان بررسی می‌گردد. این ویژگی‌ها، کاندیدهایی هستند که در طراحی یک «جداکننده[2]» به ما کمک می‌کنند.

[1] - pattern reconition
[2] - classifier

ما در تصاویر مختلف، حتی به اختلال‌هایی که در اثر تغییر در روشنایی، موقعیت ماهی بر روی نوار نقاله و حتی اختلاف‌هایی که به دلیل عملکرد الکترونیکی خود دوربین به وجود آمده است، توجه می‌کنیم. هدف طبقه‌بندی الگو این است که کلاس‌هایی را برای این الگوها در نظر بگیرد، اختلال‌ها را (که به دلیل خطا در اندازه‌گیری و نه به خاطر مدل به وجود آمده‌اند) حذف نماید و مدلی که بهترین انطباق با الگوی داده شده را دارد، انتخاب کند.

فرض کنید به تجربه دریابیم که ماهی خاردار عموماً بلندتر از ماهی سالمون است. در این صورت، می‌توان ماهی‌ها را به این روش طبقه‌بندی کرد که طول نمونه داده شده، l، از یک ماهی مجهول از یک مقدار مرزی l^* بیشتر است یا نه. مقدار l^* توسط داده‌هایی مربوط به تعدادی نمونه از هر دو گونه ماهی تخمین زده می‌شود. فرض کنید این کار را انجام داده‌ایم و نموداری به شکل ۱.۱ به دست آورده‌ایم. روشن است که طبقه‌بندی تنها بر اساس ویژگی طول بسیار ضعیف عمل می‌کند و مهم نیست که l^* چه‌گونه تخمین زده شود. نمی‌توان با اطمینان دو گونه را تنها به وسیله اندازه‌گیری طول آن‌ها از هم جدا کرد.

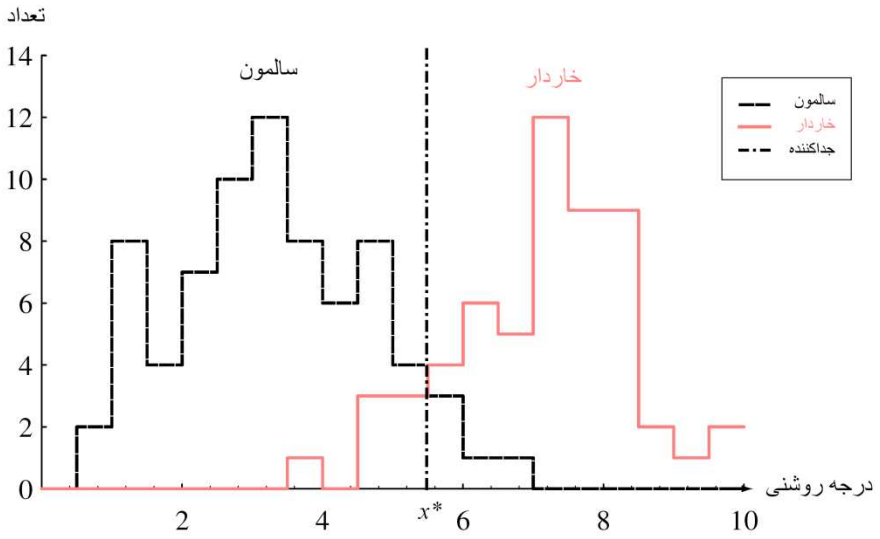


شکل ۱.۱: هیستوگرام به دست آمده برای ویژگی طول دو گونه ماهی. هیچ مقدار مرزی l^* نمی‌تواند دو گونه را بدون خطا از هم جدا کند. با استفاده از ویژگی طول به تنهایی همواره خطا داریم. مرز l^* نشان داده شده این خطا را مینیمم می‌کند.

اکنون اگر ویژگی دیگری مثل درجه روشنی پوسته، x، را نیز در نظر بگیریم، هیستوگرام به دست آمده و مقدار مرزی x^* در شکل ۲.۱، بسیار رضایت بخش‌تر از قبلی است و گونه‌ها بهتر از هم جدا شده‌اند.

در این‌جا فرض می‌کنیم که نادرست طبقه‌بندی شدن هر گونه در گونه‌ای دیگر، هزینه‌ای یکسان برای ما داشته باشد. اما الزاماً همیشه این‌طور نیست. هدف اصلی طبقه‌بندی الگو، طراحی یک «قاعده تصمیم[1]» (به عنوان مثال، تعیین مرزهای تصمیم‌گیری) است به طوری که چنین هزینه‌ای را مینیمم کند. به تجربه دریافتیم که هیچ یک از ویژگی‌ها به تنهایی نمی‌تواند جداکننده دقیقی به ما بدهد. پس برای بهبود عملکرد مدل جداکننده خود، متوسل به استفاده از چند ویژگی به طور هم‌زمان می‌شویم.

[1] - decision rule



شکل ۲.۱: هیستوگرام به دست آمده برای ویژگی درجه روشنی پوسته دو گونه ماهی . هیچ مقدار مرزی x^* نمی‌تواند دو گونه را بدون خطا از هم جدا کند. با استفاده از ویژگی درجه روشنی به تنهایی همواره خطا داریم. مرز x^* نشان داده شده این خطا را مینیمم می‌کند.

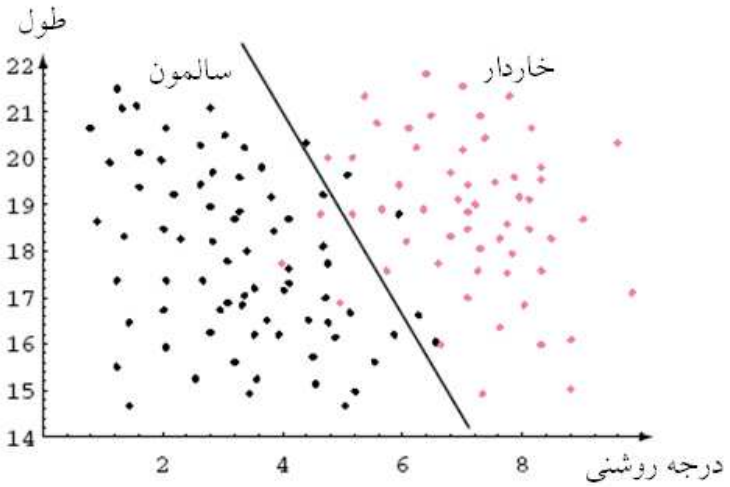
اکنون دو ویژگی طول و درجه روشنی را برای طبقه‌بندی در اختیار داریم. صرف نظر از این‌که چه طور این ویژگی‌ها اندازه‌گیری می‌شوند، داده‌ها را به بردارهای ویژگی a در یک فضای ویژگی دو بعدی تبدیل می‌کنیم یعنی $a_i = \begin{pmatrix} a_{i1} \\ a_{i2} \end{pmatrix}$ که a_{i1} ویژگی طول و a_{i2} ویژگی درجه روشنی مربوط به نمونه iام را نشان می‌دهند. این نقاط «داده‌های آموزش[1]» نامیده می‌شوند. اکنون مسأله ما تبدیل به جدا کردن فضای ویژگی‌ها به دو ناحیه می‌شود. نقاطی که در یکی از این نواحی قرار دارند، ماهی خاردار و نقاط واقع در ناحیه‌ای دیگر سالمون نامیده می‌شود. نقاط به دست آمده را در فضای ویژگی‌ها به صورت شکل ۳.۱ نمایش داده‌ایم. قاعده تصمیم ما به شرح زیر است: ماهی را به عنوان گونه خاردار می‌شناسیم اگر بردار ویژگی آن بالای «مرز تصمیم[2]» قرار گیرد و در غیر این صورت، آن را از گونه سالمون می‌شناسیم.

به نظر می‌رسد که وارد کردن ویژگی‌های بیشتر به مدل، جداکننده بهتری به ما می‌دهد. در کنار طول و درجه روشنی رنگ پوسته ممکن است ویژگی‌هایی از قبیل زاویه باله پشتی و یا محل قرار گرفتن چشم را نیز در نظر بگیریم. اما چه‌گونه بدانیم که کدام ویژگی‌ها بهترین کارکرد را دارند. برخی از ویژگی‌ها ممکن است زائد یا اندازه‌گیری آن‌ها بسیار گران باشد یا بهبود اندکی در مدل به وجود آورند و یا حتی باعث تنزل دقت مدل شوند.

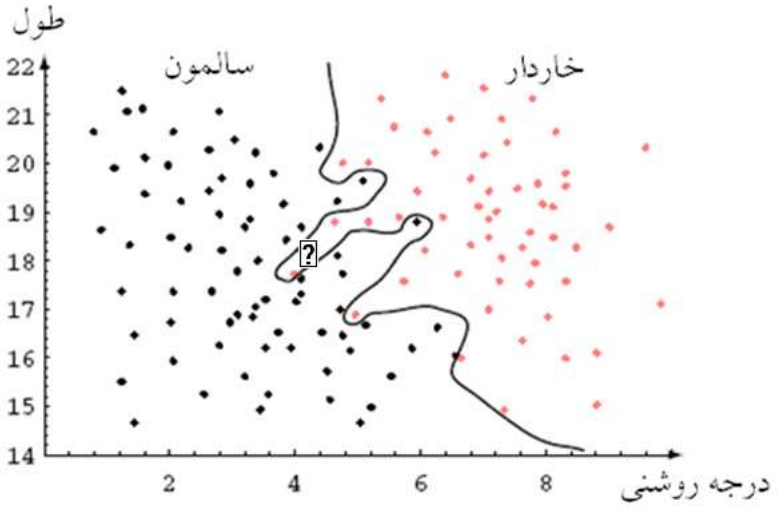
می‌توان دقت مدل را بر روی داده‌های آموزش بالا برد و مرزهای بسیار پیچیده‌تری به وجود آورد. در این صورت تمام داده‌های آموزش به طور کامل و به درستی از یکدیگر جدا می‌شوند (شکل ۴.۱ را ملاحظه کنید). اما چنین مدل جداکننده‌ای در عمل بسیار ضعیف عمل خواهد کرد. زیرا هدف اصلی از طراحی یک جداکننده، تولید مرزی است که داده‌های ناشناخته را تفکیک کند و به اصطلاح تعمیم‌پذیر باشد. مدلی که در شکل ۴.۱ به دست آمده است، منحصراً برای یک مجموعه خاص (داده‌های آموزش) خوب

[1] - training data
[2] - decision boundary

عمل می‌کند و تعمیم‌پذیری خوبی بر روی انواع داده‌ها به دست نمی‌دهد. نمونه جدیدی که در شکل با علامت ? مشخص شده است به احتمال زیاد از گونه سالمون است، اما مرز پیچیده به دست آمده، آن را در کلاس خاردار قرار می‌دهد.

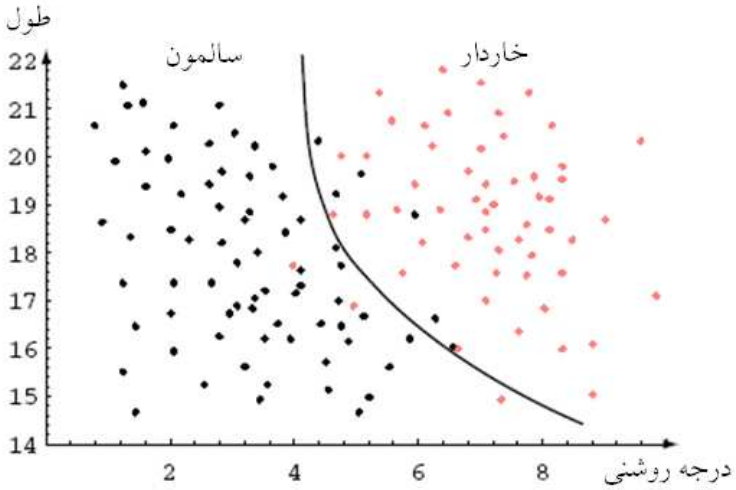


شکل۳.۱: دو ویژگی درجه روشنی و طول برای ماهی خاردار و سالمون. خط پررنگ، «مرز تصمیمِ» تعیین شده توسط جداکننده را نشان می‌دهد. مقدار خطای طبقه‌بندی کمتر از حالتی است که در آن یک ویژگی به کار رود. اما هنوز چند داده نادرست طبقه‌بندی شده وجود دارد.



شکل۴.۱: مدل‌های بیش از اندازه پیچیده مرزهای تصمیمی با پیچیدگی‌های بالا به دست می‌دهند. اگرچه، چنین مرزهایی داده‌های آموزش را به طور کامل و بدون خطا طبقه‌بندی می‌کند، اما این جداکننده بر روی داده‌های ناشناخته عملکرد ضعیفی خواهد داشت. نمونه جدیدی که در تصویر با علامت ? مشخص شده است به احتمال زیاد از گونه سالمون است، اما مرز پیچیده به دست آمده، آن را در کلاس خاردار قرار می‌دهد.

به طور طبیعی، یک روش برای به دست آوردن تخمین بهتر برای جداکننده، افزایش تعداد داده‌های آموزش است. با این وجود، در برخی از مسائل شناسایی الگو، حتی با داشتن تعداد زیادی نمونه (مانند شکل ۴.۱) باز هم جداکننده پیچیده‌تری به دست می‌آوریم که بعید است بر روی داده‌های جدید به خوبی عمل کند.



شکل۵.۱: مرز تصمیم مشخص شده، بده-بستان بهینه میان عملکرد مدل بر روی مجموعه آموزش و سادگی جداکننده را نشان می‌دهد.

اما اگر مرزهای پیچیده، خوب نیستند با چه معیاری به جداکننده‌های ساده‌تر امتیاز دهیم؟ چه‌گونه مدل طراحی شده به صورت خودکار، منحنی ساده شکل ۵.۱ را به جداکننده خطی ساده‌تر شکل ۳.۱ و مرز پیچیده‌تر شکل ۴.۱ ترجیح می‌دهد؟

فرض کنید که ما به نحوی بتوانیم این « بده-بستان[1]» ها را توسط معیارهایی بهینه کنیم. آیا می‌توانیم پیش‌بینی کنیم که مدل به دست آمده تا چه اندازه تعمیم‌پذیر برای الگوهای جدید است؟ این ها برخی از مشکلات اساسی شناسایی الگو هستند .

در مثال ماهی‌ها، لازم است که ویژگی‌ها را با دقت انتخاب کنیم تا مدلی (مثل شکل ۵.۱) به دست آوریم که قابل اطمینان باشد. یکی از جنبه‌های مهم در هر مسأله طبقه‌بندی، به دست آوردن چنین مدلی است که روابط ساختاری میان مؤلفه‌ها را در فضای ویژگی‌ها به خوبی آشکار کند و مدل مجهول واقعی را نمایان سازد. برخی از مدل‌ها به صورت بردارهایی با مؤلفه‌های حقیقی، برخی دیگر به صورت لیست مرتب شده‌ای از ویژگی‌ها و برخی به شکل بخش‌ها و ارتباط بین آن‌هاست.

ما به دنبال مدلی هستیم که در آن الگوهایی با نتیجه (کلاس) یکسان نزدیک به یکدیگر، و الگوهایی با دو نتیجه مختلف دور از هم باشند. معیاری که برای یک مدل تعریف می‌کنیم و به وسیله آن نزدیکی و دوری را می‌سنجیم قوت یا ضعف مدل را تعیین می‌کند. ممکن است بخواهیم از تعداد ویژگی‌های اندکی استفاده کنیم تا هم مرزهای ساده‌تر و هم جداکننده‌ای را که آموزش آن آسان‌تر باشد، به دست دهد. همچنین ویژگی‌هایی باشند که نسبت به اختلال‌ها[2] و خطاهای احتمالی مقاوم‌ترند. در عمل به

[1] - trade-off
[2] - noise

جداکننده‌هایی نیاز داریم که سریع‌تر کار می‌کنند و یا اجزای الکترونیکی کمتری را به کار می‌گیرند و نیاز به حافظه و مراحل پردازش کمتری دارند [۲۸] .

۳.۱ زمینه‌های مرتبط

سه زمینه‌ای که ارتباطی تنگاتنگ با شناسایی الگو دارند و در اغلب تحقیقات در این زمینه به کارگرفته می‌شوند، رگرسیون، درون‌یابی و تخمین چگالی هستند. این مسائل اغلب به طور صریح یا ضمنی به عنوان گام اول در شناسایی الگو استفاده می‌شوند.

در رگرسیون به دنبال یافتن تابعی از داده‌ها با هدف پیش‌بینی مقادیر برای یک داده ورودی جدید هستیم. رگرسیون خطی که در آن تابع، خطی است، رایج‌ترین فرم رگرسیون بوده و تاکنون بسیار مورد مطالعه قرار گرفته است. به عنوان مثال، ممکن است حس کنیم که طول ماهی سالمون مثال بالا به صورت خطی با سن و یا عرض ماهی تغییر می‌کند. در این صورت اندازه‌گیری‌هایی از سن و طول چند نمونه سالمون را جمع‌آوری کرده و از رگرسیون خطی برای پیدا کردن ضرایب استفاده می‌کنیم.

در درون‌یابی، تابع مرتبط با تعدادی داده را می‌دانیم و یا به راحتی استنباط می‌کنیم. مسأله به دست آوردن تابعی برای مقادیر میانی داده‌ها است. در مثال دسته‌بندی ماهی‌ها، فرض کنید می‌دانیم که طول ماهی نسبت به سن آن در دو هفته اول و بعد از سال دوم چه‌گونه تغییر می‌کند. در این حالت، با استفاده از یک روش درون‌یابی، ارتباط طول ماهی به سن آن را در بازه میانی دو هفته تا دو سال به دست می‌آوریم.

تخمین چگالی در مسأله طبقه‌بندی، برای تخمین زدن این احتمال که عضو یا اعضای یک گروه خاص، ویژگی خاصی دارد یا خیر مورد استفاده قرار می‌گیرد. روش‌های گوناگونی برای تخمین چگالی گروه‌ها به کار می‌رود. پس از تخمین چگالی، یک داده در کلاسی طبقه‌بندی می‌شود که بیشترین احتمال عضویت در آن را داشته باشد.

در آزمون فرض آماری، داده‌ها برای قبول یا رد یک فرض، مورد استفاده قرار می‌گیرند. به این صورت که اگر احتمال وقوع داده‌ای با در نظر گرفتن یک فرض صفر[1] از مرز تعیین شده‌ای عبور کند، فرض صفر را رد و فرض مقابل[2] را قبول می‌کنیم. در طبقه‌بندی الگوها به دنبال محتمل‌ترین فرض در میان مجموعه فرض‌ها هستیم.

در پردازش تصویر، داده‌های ورودی یک تصویر، و خروجی نیز یک تصویر است. مراحل پردازش تصویر در شناسایی الگو اغلب شامل دوران، افزایش کنتراست و تبدیلات دیگری است که اطلاعات اصلی اولیه را حفظ کنند[۲۸].

[1] - null hypothesis
[2] - alternative hypothesis

۴.۱ برخی از زیر مسأله‌های طبقه‌بندی الگو

۱.۴.۱ استخراج ویژگی

استخراج ویژگی[1] رابطه تنگاتنگی با طبقه‌بندی الگو دارد. یک مدل استخراج ویژگی ایده‌آل، عمل طبقه‌بندی را به یک فرآیند بدیهی تبدیل می‌کند. به عکس، یک جداکننده قدرتمند، به مدل‌های پیچیده استخراج ویژگی نیاز ندارد. تمایز قائل شدن میان این دو مفهوم، بیشتر به لحاظ کاربردی اهمیت دارد تا نظری.

هدف از استخراج ویژگی شناسایی مقادیری است که برای داده‌های هم‌کلاس، بسیار مشابه و برای داده‌هایی از دو کلاس مختلف، بسیار متفاوت هستند. این همان ایده یافتن ویژگی‌های متمایز کننده‌ای است که تحت تبدیلات نامرتبط حفظ می‌شوند. به عنوان مثال، در طبقه‌بندی ماهی‌ها، مکان قرار گرفتن ماهی بر روی نوار نقاله نامرتبط با کلاس متناظر با گونه ماهی است. بنابراین مدل جداکننده ما نباید به مکان ماهی حساس باشد. همچنین ویژگی‌ها وابسته به دوران (طرز قرار گرفتن عمودی یا افقی) نیز نیستند.

برخی از اصول طبقه‌بندی، می‌توانند در طراحی مدل‌های استخراج ویژگی به کار روند. تکنیک‌های طبقه‌بندی می‌توانند در کاهش حساسیت مقادیر ویژگی به اختلال‌ها مفید واقع شوند. در برخی حالات، این تکنیک‌ها را می‌توان برای انتخاب ویژگی‌های با ارزش‌تر از میان مجموعه بزرگی ازکاندیدهای ویژگی استفاده کرد.

۲.۴.۱ اختلال

تصاویر گرفته شده از ماهی‌های متعلق به یک گونه، ممکن است به دلیل لرزش نوار نقاله، تغییر در نوردهی و ایجاد سایه‌ها، تفاوت‌های زیادی داشته باشند. ما واژه اختلال را بسیار کلی تعریف می‌کنیم. هر خصوصیتی از مشاهدات داده‌شده که از الگوی واقعی برگرفته نشده و به دلیل پدیده‌های تصادفی و یا خطای وسیله اندازه‌گیری به وجود آمده باشد را اختلال می‌گویند. تمام مسائل شناسایی الگو، شامل گونه‌هایی از چنین اختلال‌هایی هستند، پرسش مهم این است: «بهترین راه برای ساختن جداکننده‌ای که به این مشکل غلبه کند چیست؟»

۳.۴.۱ بیش برازش

برای ساختن جداکننده شکل ۴.۱ در مثال ماهی‌ها مطمئناً از تابع پیچیده‌تری استفاده شده است. همان‌طور که قبلاً اشاره شد، با این‌که یک دستگاه بیش از اندازه پیچیده، ممکن است به طور کامل و بدون خطا داده‌های آموزش را از هم جدا کند، اما بعید است چنین دستگاهی بر روی داده‌های جدید و ناشناخته خوب عمل کند. این حالت را اصطلاحاً « بیش برازش[2]» می‌نامند.

درجه سختی یک مدل طبقه‌بندی، به تغییرات مقادیر بردارهای ویژگی میان اعضای هم‌کلاس، در مقایسه با اختلاف بین ویژگی‌های اعضای دو کلاس مختلف، بستگی دارد. تغییرات زیاد مقادیر ویژگی بردارهای هم‌کلاس، ممکن است به علت وجود اختلال باشد.

---

[1] - feature extraction

[2] - overfitting