طبقه بندی اجرام سماوی (کهکشان ها، ستاره ها و اختروش ها) با استفاده از داده های SDSS و تکنیک KNeighbors :

سلحشور گولان ، حسین[1] ؛

[1]دانشکده فیزیك دانشگاه تهران ، انتهاي خیابان کارگر شمالي ، تهران

چکیده

SDSS نام پروژه ای است که در آن یک تلسکوپ عظیم الجثه در حال جمع آوری داده های دقیق در باندهای طول موجی مختلف است. در این تحقیق از داده های نورسنجی این تلسکوپ استفاده کرده و با استفاده از روش همسایه یابی در یادگیری ماشین و کتابخانه ی SKlearn به جداسازی کهکشان ها، ستاره ها و اختروش ها از یکدیگر پرداخته ایم.

واژه های کلیدی: *طبقه بندی، اجرام سماوی، SDSS*

Classification of celestial objects (galaxies, stars and quasars) using SDSS data and the KNeighbors technique:

Salahshoor Gavalan, Hossein[1];

[1] Department of Physics, University of Tehran, Tehran

Abstract

SDSS is the name of a project in which a giant telescope is collecting accurate data in different wavelength bands. In this study, we used the telescope's photometric data and separated galaxies, stars, and quasars from each other using the Machine Learning Neighborhood method and SKlearn library.

Keywords*: classification,celestial objects,SDSS*

PACS No. 97

مقدمه:

آسمان محل تولد و مرگ هزاران هزار ستاره است. این ستاره ها، در مراحل مختلف خود به اجرام سماوی مختلفی تبدیل شده و خواص متفاوتی را بهمراه خواهند داشت. در این بین تشخیص ماهیت این اشیا نورانی از یکدیگر و همچنین تمیز دادن آن از کهکشان های دوردست یک وظیفه ی چالشی برای منجمان و اخترفیزیکدانان است که گاهی اوقات به رصدهای طولانی مدت و متمرکز نیاز است. این در حالی است که داده های ما از این اجرام روز بروز در حال افزایش بوده و این تفکیک بصورت دستی عملا غیر ممکن خواهد شد [1]. در این تحقیق، از آخرین داده های اپتیکی صد هزار شی سماوی که توسط SDSS جمع آوری شده (به اختصار DR17 [2]) استفاده گردیده است. جداسازی این اجرام با استفاده از زبان برنامه نویسی پایتون و کتابخانه SKlearn انجام شده است.

آشنایی با پروژه ی SDSS :

نقشه برداری دیجیتالی آسمان نام پروژه ای است که در آن یک تلسکوپ با طول قطر آینه ی 2.5 متر به جمع آوری دقیق ترین نقشه های سه بعدی کیهان در دو حوزه ی طیف سنجی و نورسنجی می پردازد [3].
این تلسکوپ، برای جمع آوری اطلاعات نوری خود از پنج باند طول موجی مختلف بهره می برد [4].

بانک داده ی SkyServer :

برای دسترسی به داده های پروژه ی SDSS، رسمی ترین درگاه مربوط به بانک داده ی SkyServer می باشد [5]. در این درگاه، ما با استفاده از زبان جستجوی SQL داده های مورد نظر خود را که شامل اطلاعات مربوط به قدر ظاهری در پنج باند طول موجی مختلف برای صد هزار کهکشان و اختروش است را جمع آوری کردیم.

```
SELECT TOP 100000
p.objid,p.ra,p.dec,p.u,p.g,p.r,p.i,p.z,
s.specobjid, s.class
FROM PhotoObj AS p
JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
  p.u BETWEEN 0 AND 19.6
  AND g BETWEEN 0 AND 20
```

شکل (1) : کد دستوری برای جمع آوری داده از پایگاه SkyServer

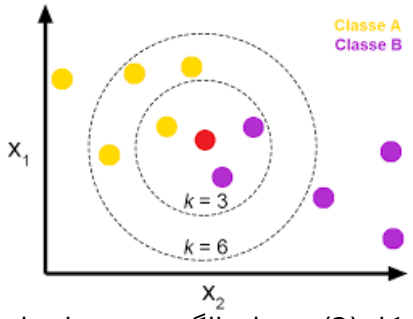
همچنین درصد فراوانی هرکدام از این اجرام در مجموعه داده ی تهیه شده در جدول (1) آورده شده است:

جدول (1) : درصد فراوانی هریک از اجرام آسمانی در داده ها

Galaxy	50.991
Star	38.344
Quasar	10.665

استفاده از کلاس KNeighbors در کتابخانه ی SKlearn برای طبقه بندی داده ها:

از الگوریتم های مهم و سودمن طبقه بندی برای داده های برچسب زده شده در ماشین لرنینگ تکنیک همسایه یابی است [6]. در این روش، برنامه به تعداد k همسایه انتخاب کرده و بر اساس انتخاب کاربر، به آن ها وزن یکسان یا برحسب فاصله می دهد (بقیه وزن صفر می گیرند)، در نهایت با استفاده از برچسب داده ها، به کمیت مجهول یک شناسه ی خاص را نسبت می دهد. انتخاب تعداد همسایه های مورد نیاز برای بررسی کاملا به داده ی مورد نظر بستگی داشته و از طریق آزمون و خطا تعیین شده و قانون کلی خاصی را پیروی نمی کند. برای مثال در شکل (2)، بسته به تعداد همسایه های انتخابی، نقطه ی مجهول کلاس های مختلفی را می پذیرد:



شکل (2) : منطق الگوریتم همسایه یابی

مروری بر تحقیقات مرتبط گذشته:

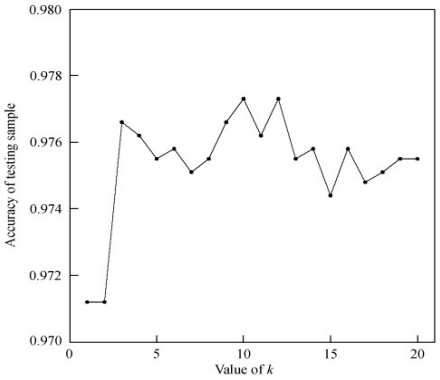
در تحقیق انجام شده در [7] ، از داده های SDSS برای طبقه بندی اجرام سماوی استفاده شده با این تفاوت که تعداد ویژگی ها هفده و تعدا کل داده ها هزار عدد است. همچنین محققان این مقاله از چندین روش ماشین لرنینگ برای آموزش استفاده کرده اند که در جدول (2) به الگوریتم استفاده شده و نتیجه ی آن پرداخته ایم:

جدول (2) : الگوریتم های یادگیری ماشین استفاده شده در تحقیق [7] به همراه نتایج آن

Name	test_acc	train_acc
Logistic Regression	0.796	0.796
SVM	0.647	0.726
Random Forest	0.98	1
Decision Tree	0.98	0.99

در تحقیق انجام شده در [8] ، از الگوریتم مشابه با این مقاله برای تمیز دادن هسته ی کهکشان های فعال از ستاره ها و کهکشان های معمولی استفاده شده است. برای این منظور از داده های اپتیکی، فروسرخ و

پرتو ایکس این اشیا سماوی استفاده شده است. طبق نتایج آن ها، روش همسـایه یـابی در این کلاس بنـدی بسیار موفق عمل کـرده و از این رو ما نـیز این رویکـرد را اتخـاذ کـردیم. نتـایج این تحقیق بـرای مجموعه ی تست بر حسب تعداد همسایه ها در شکل (3) آمده است:



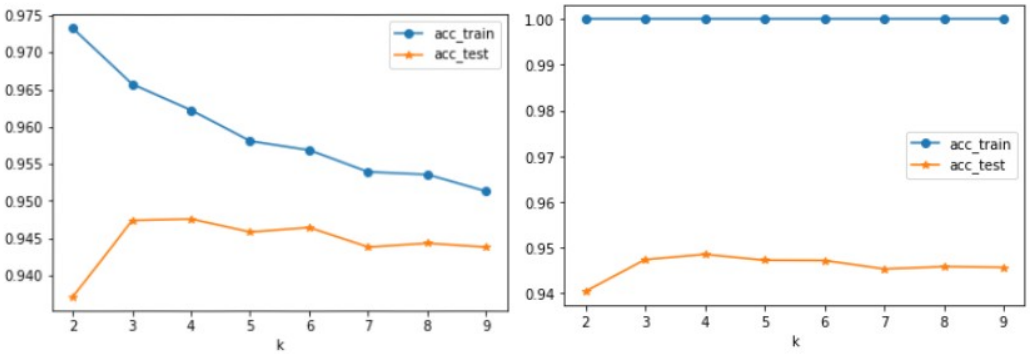
شکل (3) : دقت الگوریتم همسایه یابی برای تحقیق [8] در مجموعه ی تست بر حسب تعداد همسایه های مختلف

پیش پردازش داده ها:

در این تحقیق داده ی اصلی به دو مجموعه ی آموزش و تست تقسیم شـده به طـوری که هشـتاد در صد داده ها (3200 داده) به مجموعه ی آموزش و باقی به تست اختصاص داده شده است.

اجرای فرآیند یادگیری با تابع وزن های مختلف:

اگر تابع وزن را در دو مد uniform و distance انتخاب کنیم، دقت نتایج برای دو مجموعه ی آموزش و تست با تعداد همسایه های مختلف در شکل (4) گزارش شده است:



شکل (4) : دقت نتایج برای دو مجموعه ی آموزش و تست در k های مختلف در حالتی که تابع وزن یکنواخت (چپ) یا فاصله (راست) انتخاب شده است

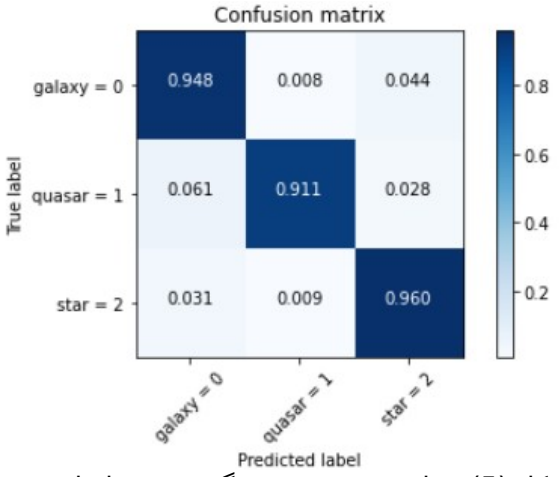
نتیجه گیری:

در مقایسه با تحقیق [7] ، ما بسیاری از ویژگی ها را حذف و تعداد داده ها را افزایش دادیم. دقت نتایج ما بسیار بالا بوده و همچنین خطر بیش برازش در آن نسبت به تحقیق ذکر شده بسیار کمتر است.

طبق نتایج دو تحقیق اشاره شده، می توان استنباط کرد که دو روش درخت تصمیم و همسایه یابی برای کلاس بندی داده های اخترفیزیکی عملکرد بالایی دارند.

در تحقیق ما، اگرچه بین دو تابع وزن مختلف تفـاوت چنـدانی مشـاهده نمی شـود، اما مد فاصـله از مد یکنواخت با حاشیه ی کمی عملکرد بهتری را نشان می دهد. همچنین در مد یکنواخت دقت مجموعه ی تست در تعداد سه همسایه، و در مد فاصله در تعداد چهار همسایه بیشترین مقدار است.

برای بررسی بیشتر، ماتریس درهم ریختگی برای مد فاصله و تعداد چهـار همسـایه در شـکل (5) رسم شده است:



شکل (5) : ماتریس در هم ریختگی در مد فاصله و $k = 4$

مرجع‌ها:

[1] : Scargle, Jeffrey D., Kamal M. Ali, and Ashok N. Srivastava. *Advances in machine learning and data mining for astronomy*. Ed. Michael J. Way. Boca Raton, FL: CRC Press, 2012.

[2] : Accetta, Katherine, et al. "*The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar and APOGEE-2 Data*." arXiv preprint arXiv:2112.02026 (2021).

[3] : **https://www.sdss.org/**

[4] : Fukugita, M., et al. "*The Sloan digital sky survey photometric system*." *The Astronomical Journal* 111 (1996): 1748.

[5] : **http://skyserver.sdss.org/dr17/SearchTools/sql**

[6] : Zhang, Zhongheng. "*Introduction to machine learning: k-nearest neighbors*." *Annals of translational medicine* 4.11 (2016).

[7] : Sharma, Siddharth, and Ruchi Sharma. "*Classification of Astronomical Objects Using Various Machine Learning Techniques*." *Advances in Data Sciences, Security and Applications*. Springer, Singapore, 2020. 275-283.

[8] : Li, LiLi, YanXia Zhang, and YongHeng Zhao. "*k-Nearest Neighbors for automated classification of celestial objects*." *Science in China Series G: Physics, Mechanics and Astronomy* 51.7 (2008): 916-922.