Machine Learning Course

Assignment #5:

Random Histogram Forest for Unsupervised Anomaly Detection

آرمین خیاطی
9931153

پویان انصاری راد
9960565

Summer 2021

فهرست

مقدمه

در این تمرین مقاله زیر مورد بررسی و پیاده سازی قرار گرفته است :

Random Histogram Forest for Unsupervised Anomaly Detection

Published in: 2020 IEEE International Conference on Data Mining (ICDM)

Date of Conference: 17-20 Nov. 2020

Date Added to IEEE Xplore: 09 February 2021

ISBN Information:

ISSN Information:

INSPEC Accession Number: 20424249

DOI: 10.1109/ICDM50108.2020.00154

Publisher: IEEE

Conference Location: Sorrento, Italy

موضوع مقاله در ارتباط با شناسایی داده های غیر نرمال یا Outlier بصورت غیر نظارت شده توسط روش Random Histogram Forest یا به اختصار RHF میباشد. به طور کلی ، تشخیص ناهنجاری شامل موارد زیر است:

شناسایی داده هایی که مقدار ویژگی ها یا فیچر های آنها به طور قابل توجهی از بقیه داده های ورودی فاصله دارد. این چالش یکی از گسترده ترین مشکلات مورد مطالعه در یادگیری ماشین بوده و دارای کاربردهای متنوعی از جمله در تشخیص نفوذ در شبکه، تشخیص های پزشکی و ... میباشد. روش های مختلف و گوناگونی در سالهای اخیر مورد بررسی قرار گرفته و ارایه شده است با این حال همچنان راه کار کاملا رضایت بخشی یافت نشده است. روش جنگل هیستوگرام تصادفی یا همان RHF یک روش مفید بدون نظارت است. ثابت شده است که رویکرد احتمالاتی تشخیص ناهنجاری در شناسایی ناهنجاری ها بسیار موثر است. در این روش از fourth central moment (معروف به kurtosis) استفاده می شود.

رویکرد ارایه شده در مقاله بر پایه ساخت Random Forest بر پایه تمام نمونه های ورودی است. الگوریتم پیشنهاد شده دارای مرتبه زمانی خطی براساس تعداد نمونه های دیتاست میباشد. ویژگی دیگر الگوریتم استفاده از امتیاز دهی بر مبنای محاسبه fourth central moment (معروف به kurtosis) برای هدایت فرآیند جستجوی نمونه های ناهنجار یا آنومال میباشد.

نکته دیگر قابل ذکر ارزیابی الگوریتم های مورد بررسی و الگوریتم ارایه شده توسط معیار Average Precision یا AP یا همان اندازه ناحیه زیر نمودار precision-recall است. همانطور که در مقاله ذکر شده، معیار مرسوم ROC برای ارزیابی نتیجه این الگوریتم ها در این زمینه کاری نمیتواند میزان کارایی واقعی این الگوریتم ها را به دقت اندازه گیری کند.

روش مذکور در مقاله فوق بر روی تعداد 38 دیتاست مختلف اجرا شده و با بقیه روش های موفق تاکنون مقایسه شده است. معیار مقایسه روش ها AP یا Average Precision انتخاب شده.

در این تمرین الگوریتم ارایه شده بر روی یکی از دیتاست های معرفی شده با عنوان Ionosphere پیاده سازی شده و با متریک معرفی شده موفقت الگوریتم ارزیابی میگردد و نتیجه با نتایج ارایه شده در مقاله مقایسه خواهد شد.

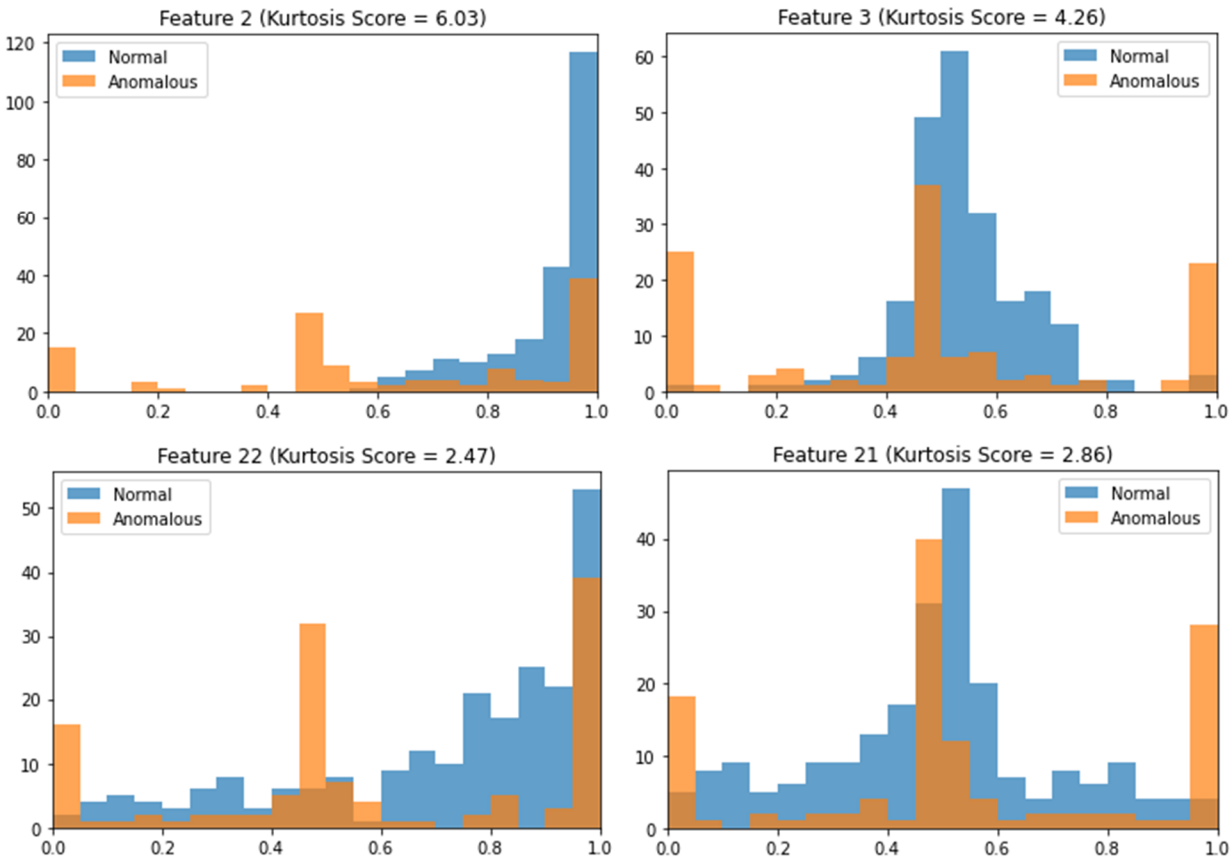
بخش اول

در این بخش دیتاست مورد اشاره بارگذاری میگردد. این دیتاست دو کلاسه است و مقدار 0 نشان دهنده داده مطلوب یا نرمال یا Inlier بوده و مقدار 1 نشان دهنده داده ناهنجار یا غیرنرمال یا Outlier میباشد :

```
---------------------------DataSet : Ionosphere---------------------------
x_train.shape = (351, 34)
y_train.shape = (351,)
m = 351
n = 34
c = 2
y_unique = [0 1]
y = 0 is Inlier
y = 1 is Outlier
--------------------------------------------------------------------------
```

برای درک بهتر موضوع درصورتیکه نمودار تابع توزیع نرمال شده هر کدام از ویژگی ها یا Feature ها را در هر کدام از دو کلاس معرفی شده رسم کنیم درک بهتری از اینکه کدام ویژگی ها بیشتر آنومالی اطلاعات دارند بدست میآوریم. همچنین برای هر فیچر مقدار fourth central moment (معروف به kurtosis) را نیز محاسبه و نمایش میدهیم.

نمودار چند نمونه از 34 ویژگی دیتاست در زیر نمایش داده شده است :



همانطور که در تصاویر فوق ملاحظه میکنید، توزیع های رنگ نارنجی توزیع داده های یک فیچر در کلاس آنومالی است و توزیع های رنگ آبی توزیع داده های فیچر در کلاس نرمال است. همانطور که ملاحظه میکنید، نمودار داده های آنومالی یا ناهنجار به صورت دنباله دار و بوده و بعضا ناپیوسته هستند. قابل مشاهده است که فیچر هایی که Kurtosis Score بالاتری دارند، آنومالی بیشتری نیز در نمودار توزیع آنها مشاهده میشود.

بخش دوم :

روش RHF یک مدل Ensemble است که دیتاست را به n گروه مختلف تقسیم میکند. الگوریتم داده های ورودی را با انتخاب نقاط تصادفی به گروه هایی تقسیم و امتیاز یندی میکند، گروه های بزرگتر نشان دهنده میزان کمتری از احتمال آنومالی بودن دارند. با تکرار فرآیند فوق میتوان میزان آنومال بودن این نقاط را اندازه گیری کرد.

روش کار به این صورت است که از میان مجموعه فیچر ها، فیچری با توجه به Kurtosis Score آن انتخاب میشود و سپس مقدار تصادفی در بازه حداقل و حداکثر مقدار آن فیچر در داده ها انتخاب شده و سپس داده ها براساس این مقدار فیچر اسپلیت شده و درخت هیستوگرام تصادفی RHT آن ایجاد میشود.

میزان عمق درخت را خودمان براساس تجربه مشخص میکنیم. این روند را تا رسید به برگ های درخت مورد نظر ادامه میدهیم تا درخت RHT ما براساس پارامتره های انتخاب شده کامل شود. در نهایت میزان آنومالی بودن یک نمونه برابر میشود با مجموع درجه Imformation Content (aka self-entropy) همه برگ های درخت های ایجاد شده.

این فرآیند را برای تعداد قابل انتخابی درخت تکرار میکنیم و امتیاز های هر نود برگ را در تمام درخت ها با هم جمع میکنیم. در نهایت به ازای هر داده ، یک امتیاز به آن داده تعلق خواهد گرفت که آن امتیاز میزان غیر مترقبه بودن یا آنومال بودن آن داده در تمام درخت های ایجاد شده است.

نتیجه عملیات فوق، لیست امتیاز های زیر است، به هر سمپل یک امتیاز تعلق گرفته است :

```
[1.45386296 2.39378544 1.25689309 3.01196574 1.71076305 1.53897431
 2.15164112 3.39676512 1.79236842 2.0002943  2.00531194 2.7509092
 1.6062824  2.10817894 1.60790441 2.95481853 1.21735999 4.69306373
 1.58515013 4.05395743 1.45665955 3.92596173 1.10415578 3.88515693
 1.35492017 2.94195076 2.1164016  4.28061164 2.35235613 4.55731487
 1.80381929 2.90987201 1.62772015 2.23253945 1.84735868 2.63054337
 1.28493764 3.85074782 1.18267052 1.69419136 1.73738612 4.11592592
 1.74732828 3.40487302 1.71718726 3.3576292  1.48036584 2.71852748
 1.27975318 2.62591695 1.61541315 3.77798014 3.49733822 4.18875981
 1.47240361 4.01960374 1.63749518 4.23894725 1.44143228 3.22542026
 1.46448636 2.3167101  1.67026382 3.09856301 1.56189888 1.88283379
 1.31132569 3.2578852  2.72744285 2.92288294 1.76192646 4.06187387
 1.90431052 1.83236521 1.69619903 3.79417788 1.66848259 4.09981549
 3.0321198  3.81452937 2.10627933 3.22530507 1.78583991 1.72228758
 1.23347554 1.70720715 1.24212686 1.86004977 1.28129222 2.12632978
 1.67487626 2.65144951 1.36400015 2.33315948 1.2863299  2.47347212
 2.36301819 2.49593068 3.62671652 1.19663932 3.10180589 1.38312358
 2.01648324 1.46223129 1.96061751 1.75430167 1.7945323  1.27031412
 2.14328036 2.47308578 2.13403201 2.29122769 2.03906815 2.52389565
 2.14503869 1.41867137 1.28019849 1.26353255 2.63137203 1.2950013
 1.75066421 1.97143487 2.53180584 1.3673462  2.42860249 1.76075985
 2.13281943 1.24918865 2.44490371 1.23008163 1.90538825 1.34060292
 2.67265917 2.06905588 2.40600627 2.42600617 3.63719885 1.427301
 1.72664437 1.52785347 3.28977522 1.20942918 2.33264592 1.51640513
 1.39395266 1.50329227 1.95613511 1.28209488 1.93332411 1.20651825
 2.13347307 1.26134035 1.44518931 2.11411975 1.50075371 1.08125337
 1.44054066 1.97269572 1.49455786 2.41355522 1.44742759 1.16953481
 4.02717855 2.37941768 2.93448631 2.84316321 3.94409023 3.15470515
```

1.9734164 1.61951103 3.8885653 1.24938857 3.0377437 1.5684094
2.04844093 1.53899273 3.09422977 1.27724702 1.54514782 1.93244219
2.86314758 1.60694044 4.15047509 1.93391866 3.8114869 2.35929556
3.44277669 3.62424032 4.09497469 2.23036878 3.150998   1.70853323
2.54972614 1.49038018 4.1097641 1.40892186 2.9369042 1.52716593
3.15230551 1.45143767 3.22263573 1.9693241 3.51437793 1.25705103
3.74920992 1.27581558 4.22670666 1.40681563 2.29820893 1.80485177
3.82670018 1.33069794 2.79057681 1.63816342 3.28322495 2.31385863
3.2100512 1.6536229 3.17803998 2.17206734 4.3115166 1.83210941
3.10772587 1.8842294 2.49811753 2.13616628 3.13404187 2.228554
3.31932992 1.81846072 2.00071884 1.18835283 3.76473432 1.24074045
1.97201127 1.7512057 1.80805954 1.19485336 1.57170176 1.15796497
2.10433585 1.16503687 2.00977348 1.12900568 1.96824783 1.12256587
1.90057673 1.26260919 2.01648324 1.98681243 2.96528543 2.2375612
2.26548547 2.22832887 2.09388433 2.36577868 1.24894696 1.37039465
1.22039135 1.06851152 1.26653039 1.87130345 2.02868249 2.05639012
2.62114742 2.31223316 2.04162782 1.52794524 1.37738752 1.12189551
1.27539866 1.52912489 1.53096827 1.26378243 1.16080603 1.11727027
1.22237075 1.24154527 1.9538291 2.22949259 2.40729578 2.53231135
2.50615852 2.21545147 1.57926714 1.47142669 1.35767242 1.09645787
1.19505109 1.16878838 1.40165382 1.30755838 1.1758768 1.15051096
1.29173673 1.28892288 2.48765764 2.6903868 2.60210755 2.36072802
2.8012541 2.39569354 1.27468148 1.16840149 1.32265624 1.5208559
1.4047632 3.10800922 2.48850077 2.33528513 2.79863355 2.58298136
1.34342983 1.26953972 2.13181977 2.28982058 2.47375054 1.36544567
1.13063401 1.15915477 1.26509708 1.14407849 1.19797184 1.41446475
2.1410982 2.30318694 2.49945234 2.41567786 2.22207938 1.18115468
1.25070656 1.17532413 1.09091438 1.24788242 1.23967684 1.31936634
2.20818758 2.56789109 2.09661453 2.39394515 2.12865944 1.22680906
1.11990776 1.23963506 1.0789945 1.55576478 1.25746528 1.25982174
1.17750157 1.14630118 1.13535177]

نتایج :

درصورتیکه الگوریتم اشاره شده را 10 مرتبه بر روی دیتاست با پارامتر های زیر اجرا کنیم و برای ارزیابی نتایج از AP یا Average Precision استفاده کنیم میتوانیم مشاهده کنیم که در مقایسه با سایر روش ها، روش ارایه شده در مقاله مورد بررسی نتیجه مطلوب تری را بدست آورده است :

num_trees = 100

max_height = 5

run = 10

```
Run = 0 --> Average Precision Score :0.8011630656066201
Run = 1 --> Average Precision Score :0.805583084007407
Run = 2 --> Average Precision Score :0.8058630558718594
Run = 3 --> Average Precision Score :0.8137809118067669
Run = 4 --> Average Precision Score :0.797966788171437
Run = 5 --> Average Precision Score :0.804028313224868
Run = 6 --> Average Precision Score :0.8160640391211081
Run = 7 --> Average Precision Score :0.8087560484487872
Run = 8 --> Average Precision Score :0.804004628601483
Run = 9 --> Average Precision Score :0.8225148973739984

Mean of 10 Runs --> Average Precision Score :0.8079724832234335
STD of 10 Runs --> Average Precision Score :0.007065750240697475
```

در مقاله پایه میزان AP برای این دیتاست : 0.819 با انحراف معیار 0.006 اعلام شده است.

در پیاده سازی میزان AP برای این دیتاست : 0.807 با انحراف معیار 0.007 بدست آمده است.

	n	d	anomalous - duplicates	HBOS	PPCA	OCSVM	KNN	KthNN	LOF	ISO	RHF_K	RHF_R
musk	3060	166	3.1% - 0%	0.904	**1.0**	**1.0**	0.432 ± 0.023	0.626 ± 0.027	0.239 ± 0.013	0.980 ± 0.021	**0.994 ± 0.007**	0.990 ± 0.008
http_logged	567498	3	0.4% - 97%	0.242	0.769	0.492	0.009 ± 0.001	0.009 ± 0.001	0.022 ± 0.003	0.947 ± 0.033	**0.982 ± 0.002**	0.990 ± 0.001
kdd_smtp29	96554	3	0.03% - 0%	0.980	0.773	0.405	0.090 ± 0.002	0.104 ± 0.002	0.014 ± 0.001	**0.989 ± 0.001**	0.954 ± 0.008	0.970 ± 0.005
breastcancer	683	9	34.9% - 1.2%	0.878	0.958	0.918	0.933 ± 0.001	0.942 ± 0.002	0.294 ± 0.007	**0.967 ± 0.004**	0.952 ± 0.005	0.962 ± 0.003
shuttle	49097	9	7% - 0%	0.911	0.915	0.907	0.182 ± 0.001	0.188 ± 0.001	0.118 ± 0.002	**0.976 ± 0.005**	0.933 ± 0.006	0.951 ± 0.003
satimages	5803	36	1.2% - 0.03%	0.732	0.872	0.965	0.443 ± 0.013	0.554 ± 0.033	0.028 ± 0.002	0.923 ± 0.006	**0.926 ± 0.008**	0.918 ± 0.010
kdd_ftp	5214	3	26.7% - 79.7%	0.391	0.846	0.596	0.259 ± 0.001	0.261 ± 0.001	0.284 ± 0.002	0.428 ± 0.014	**0.922 ± 0.008**	0.909 ± 0.007
ionosphere	351	33	35.8% - 0.7%	0.28	0.747	0.839	**0.922 ± 0.003**	0.866 ± 0.009	0.851 ± 0.004	0.812 ± 0.005	**0.819 ± 0.006**	0.800 ± 0.006
kdd99G	620098	29	0.17% - 1.33%	0.585	0.683	0.325	0.177 ± 0.003	0.190 ± 0.003	0.004 ± 0.001	0.531 ± 0.002	**0.774 ± 0.056**	0.577 ± 0.033
kdd_http29	623091	29	0.64% - 31.2%	0.536	0.758	0.499	0.096 ± 0.002	0.108 ± 0.001	0.017 ± 0.003	0.537 ± 0.012	**0.770 ± 0.076**	0.501 ± 0.015
kdd_http_distinct	222027	3	0.03% - 0%	0.049	0.637	0.373	0.352 ± 0.010	0.375 ± 0.007	0.027 ± 0.001	0.017 ± 0.005	**0.743 ± 0.042**	0.795 ± 0.015
mulcross	262144	4	10% - 0%	0.064	**0.979**	0.643	0.052 ± 0.001	0.052 ± 0.001	0.171 ± 0.001	0.565 ± 0.034	**0.733 ± 0.032**	0.730 ± 0.041
satellite	5100	36	1.4% - 0%	0.500	0.583	0.622	0.563 ± 0.012	0.612 ± 0.006	0.187 ± 0.001	0.639 ± 0.014	**0.651 ± 0.015**	0.650 ± 0.015
magicgamma	19020	10	35.1% - 1.7%	0.467	0.586	0.626	**0.735 ± 0.001**	0.728 ± 0.001	0.540 ± 0.003	**0.648 ± 0.008**	0.624 ± 0.010	0.626 ± 0.012
wbc	378	10	5.5% - 4.7%	**0.699**	0.556	0.529	0.546 ± 0.001	0.554 ± 0.002	0.573 ± 0.011	**0.591 ± 0.026**	0.577 ± 0.013	0.612 ± 0.011
cardio	1831	31	9.6% - 0.56%	0.416	**0.612**	0.533	0.363 ± 0.003	0.384 ± 0.002	0.156 ± 0.001	0.557 ± 0.027	**0.567 ± 0.023**	0.553 ± 0.023
penglobal	809	16	11.1% - 0%	0.237	0.301	0.569	**0.897 ± 0.003**	0.864 ± 0.044	0.566 ± 0.018	**0.612 ± 0.027**	0.556 ± 0.039	0.553 ± 0.043
kdd_http	623091	3	0.64% - 98.1%	0.204	**0.550**	0.369	0.010 ± 0.001	0.010 ± 0.001	0.017 ± 0.002	0.488 ± 0.049	**0.550 ± 0.009**	0.572 ± 0.002