سحر محمدی 9812762305

فاز دوم پروژه هوش محاسباتی

1. در قدم اول میخواهیم از بین 15 ویژگی که پیدا کرده ایم، ویژگی های بهتر را انتخاب کنیم.

روش اول: با هر ویژگی به تنهایی یک درخت تصمیم بسازیم و طبق پارامتر دقت، ویژگی های برتر را پیدا میکنیم.

1. میانگین 78%
2. انحراف از معیار 95%
3. بیشترین مقدار یا max 95%
4. min 95%
5. skewness 85%
6. کشیدگی یا kurtosis 78%
7. ptp 94%
8. Twopp 73%
9. Pps 77%
10. RMS یا root mean square 97%
11. Margin factor 93%
12. Shape factor 94%
13. Impulse factor 75%
14. Third freq 74%
15. Forth freq 76%

اگر بخواهیم دقت 76 به بالا را انتخاب کنیم در نتیجه انهایی که انتخاب میشوند: mean, std, max, min, skew, kurt, ptp, rms, margin, shape, 4f

1. اما ممکن است برخی از این ویژگی ها کورلیشن بالایی داشته باشند و استفاده از هر دوی انها اطلاعات جدید و بدردبخوری به ما ندهند. پس در قدم بعدی میخواهیم ویژگی هایی که کمترین کورلیشن را باهم دارند انتخاب کنیم.

ابتدا کورلیشن دوبه‌دوی همه ویژگی ها را محاسبه میکنیم، قدر مطلق میگیریم و مرتب میکنیم.

1. از بین انهایی که باهم کورلیشن پایین دارند، اولین زوجی که شامل twopp نیست(چون twopp به تنهایی از دقت بالایی برخوردار نیست.) میشود mean, max

سپس std را چک میکنیم که ایا میتوان انتخاب کرد یا نه، ابتدا کورلیشن std با max را میگیریم، که مقدار نزدیک به 1 است پس این ویژگی اطلاعات جدیدی به ما نخواهد داد پس انتخاب نمیشود.

Min و ptp هم کورلیشن بالایی با max داشتند.

و اما skew، کورلیشن پایین با max دارد، حال کورلیشن آن با mean را هم بدست می اوریم، که آن هم پایین است. پس این ویژگی اطلاعات جدیدی به ما میدهد و تکراری نیست. پس انتخاب میشود.

برای انتخاب ویژگی باید کورلیشن آن را با همه ی قبلی ها چک کرد و بیشتر کورلیشن نباید مقدار زیادی باشد.

به همین ترتیب در نهایت ویژگی هایی زیر انتخاب میشوند:

Mean, max, skew, kurtosis, pps, 4freq

1. اما از یک نگاه دیگه، برای اینکه ببینیم هر ویژگی نسبت ویژگی قبلی چقد اطلاعات جدید به ما میده، لازمه یک بار درخت با ویژگی اول را رسم کنیم، بعد ببینیم برای کدام داده ها پیشبینی اشتباه صورت گرفته، ایا با اضافه کردن ویژگی دوم و تشکیل درخت جدید، کلسیفیکیشن بهتر شده؟

برای این کار، ابتدا درخت را با max رسم ساختیم، که دقت 95% دارد و برای 5 ویژگی اشتباه پیشبینی میکند.

این کار را برای std هم کردیم که دقت آن هم 95% است ولی انهایی که اشتباه کرده به جز یک مورد با max متفاوت است.

و طبق چیزی که انتظار میرفت رسم درخت بکمک دو ویژگی std و max دقت 99% داشت.

و در نهایت هیچ ویژگی دیگر دقت را از 99 بهتر نکرد. پس بنظر میرسد که ویژگی های دیگر نمیتوانند اطلاعات جدیدی بدهند و همین دو ویژگی کافی است.

1. در این مرحله میخواهیم بکمک خوشه بندی، کیفیت و سرعت کلسیفیکیشن را افزایش بدهیم. یعنی ابتدا داده های شبیه را در یک خوشه قرار دهیم، و بعد هر خوشه را جداگانه طبقه بندی کنیم.
2. تست کردن حالت های مختلف

3 حالت دو کلاسه:

ABCD E که درخت با این حالت دقت 99% دارد.

AC E 93%

D E 100%

1 حالت سه کلاسه:

AB CD E 99%

مراجع:

[1] <https://www.kaggle.com/code/oybekeraliev/time-domain-feature-extraction-methods>

[2] <https://www.kaggle.com/code/oybekeraliev/frequency-domain-feature-extraction-methods>

[3] <https://www.geeksforgeeks.org/decision-tree-implementation-python/>

[4] <https://www.geeksforgeeks.org/sort-correlation-matrix-in-python/>

[5] <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html>

[6] <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>