سوالات تمرین شماره دو BIG DATA

تاریخ تحویل : 1400/10/6 نمره : 2

1. پایگاه داده BIG2015 که توضیحات آن در یکی از مقالات پیوست آورده شده است حاوی بدافزارهایی به دو صورت کد اسمبلی و کد باینری میباشد. میخواهیم با استفاده از استخراج ویژگی فرکانس تکرار کدهای اسمبلی برای هر بد افزار و به کمک RANDOM FOREST در SPARK دقت کلاسبندی و زمان اجرا را برای ۹ کلاس اصلی بد افزار در حالات زیر بدست آوریم.

الف: دو نمودار دقت و زمان اجرا را برای تعداد درخت 10-20-30-40-50 و حداکثر عمق درخت 10 رسم کنید.

ب: مقادیر ماتریس سردرگمی و متریک های زیر را برای بهترین جواب ارایه دهید Precision, Recall, F-measure, True Positive Rate, False Positive Rate

خروجی گزارش شامل کد، نتایج، نمودار و تحلیل آن می باشد.

2. با استفاده از روشی که در مقاله دوم پیوست اورده شده است بردارهایی از تصاویر فایل باینری بدافزار برای آموزش و تست یک درخت تصمیم گیری در SPARK استفاده کنید.

الف : نمودار دقت را با تغییر عمق درخت از k=3 to8 رسم نمایید.

ب: با در نظر گرفتن عمق بهینه درخت پارامتر MaxBins را بین مقادیر 4 و 8 و 16 و 32 تغییر دهید. نمودار دقت خروجی را بر حسب این پارامتر رسم نمایید و تغییرات موجود

در دقت نتایج را توضیح دهید. گزارش شامل کد و نتایج و نمودارهای هر مرحله و تحلیل جواب ها می باشد.