تمرین و پروژه شماره ۱ درس داده کاوی، گزارش سؤال ۳

تکین جزایری، ۹۸۱۳۰۰۶ – امیرحسین رجبی، ۹۸۱۳۰۱۳

۱. برای این قسمت از WEKA Experimenter استفاده شده است؛ به این صورت که در قسمت Experiment Type، Cross-validation با ۱۰ فولد قرار داده می شود. همچنین الگوریتم J48 از قسمت Algorithms انتخاب می شود و تعداد مرتبه تکرار را برابر با ۱ قرار می دهیم. بعد از run کردن، در قسمت Analyse بعد از تعریف یک Experiment جدید، یک جدول تولید می کنیم که سطرهای آن (Rows) را Fold و ستون های آن (Cols) را Percent_incorect انتخاب می کنیم. نتیجهٔ این آزمایش در فایل Q3-1.csv موجود است که به ما نشان می دهد که قسمت های مختلف دیتاست ممکن است درصد های متفاوتی از خطا را در ساخت درخت تصمیم حاصل کند.

۲. برای این قسمت WEKA Explorer مورد استفاده قرار گرفته است. برای ایجاد منحنی یادگیری، در قسمت Classify به عنوان Classifier، FilteredClassifier انتخاب می کنیم و در تنظیمات آن، به ترتیب J48 و RemovePercentage را برای فیلدهای classifier و filter تعیین می کنیم. مدل ساخته شده در این مرحله در Q3-2 mod.model قابل مشاهده است.

برای مشاهدهٔ منحنی یادگیری، گزینهٔ Visualize threshold curve را انتخاب می کنیم و در پنجرهٔ باز شده، مقادیر X و Y را به ترتیب برابر با Sample Size و Precision تعیین می کنیم. تصویر نمودار حاصل در فایل Q3-2 graph.JPG قابل مشاهده است.

این نمودار نشان می دهد که بیشترین دقت و بیشترین میزان یادگیری الزاماً در بیشترین تعداد نمونه بدست نمی آید. به طور مثال در اینجا بیشترین دقت در با استفاده از حدود ۱۷ درصد از داده های حاصل می شود.

۳. در این قسمت با استفاده از WEKA Explorer، الگوریتم J48 را دو مرتبه روی دیتاست داده شده اجرا می‌کنیم؛ با این تفاوت که در مرتبهٔ دوم از تنظیمات Classifier، unpruned را برابر با True قرار می‌دهیم. نتیجهٔ این دو اجرا در فایل های Q3-3 mod pruned.model و Q3-3 mod unpruned.model آمده است. مشاهده می شود که در نتیجهٔ دوم، تعداد برگ ها و اندازهٔ درخت تقریباً ۱۲ و ۱۱ برابر شده اند. نتایج حاصل به صورت زیر است:

	Number of Leaves	Size of the tree
pruned	564	710
unpruned	6812	7976
Difference	**6284**	**7266**

۴. روش های انتخاب به صورت تصادفی و استفاده از بهرهٔ اطلاعات در الگوریتم های WEKA هم به ترتیب با نام های RandomTree و REPTree موجودند. مدل های ساخته شده توسط این دو الگوریتم در فایل های Q3-4 mod RandomTree.model و Q3-4 mod REPTree.model قرار داده شده است.

۵. برای اجرای الگوریتم ها به محیط Experiment می رویم. در قسمت Experiment Type حالت Train/Test Percentage Split (data randomized) را انتخاب کرده و تعداد تکرار را در قسمت Iteration control برابر با ۲۰ قرار می دهیم. همچنین در قسمت Algorithms، به ترتیب الگوریتم های RandomTree، REPTree و J48 را اضافه می کنیم. درصد جواب های درستی که از این ۶۰ مرتبه بدست آمده است در فایل Q3-5 Percent_correct.csv موجود است. این نتایج نشان می دهد که بیشترین دقت با استفاده از الگوریتم J48 (۴۱ تا ۶۰) بدست آمده است و پس از آن به ترتیب REPTree (۲۱ تا ۴۰) و RandomTree (۱ تا ۲۰) قرار گرفته اند.

همچنین جدول اندازهٔ درخت ایجاد شده در تمامی اجراها در فایل Q3-5 Serialized_Model_Size.csv آمده است و هیستوگرام برای ۲۰ اجرای اول (RandomTree) به صورت زیر است: