"بسمه تعالی"

گزارش کار پروژه نهایی

"پیش بینی بیماری ها از روی داده های میکرو آرایه های ژنتیک"

درس : داده کاوی

استاد : جناب آقای دکتر احمدی

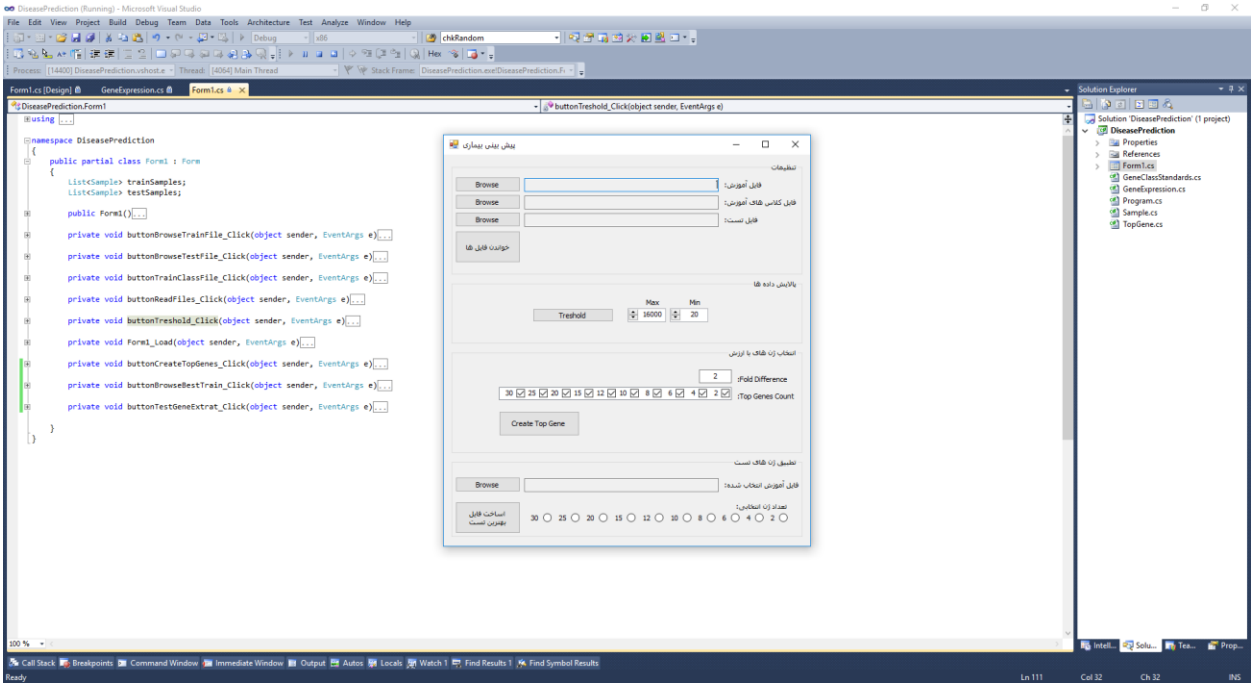
تهیه کننده : فرخنده زینالی آق قلعه

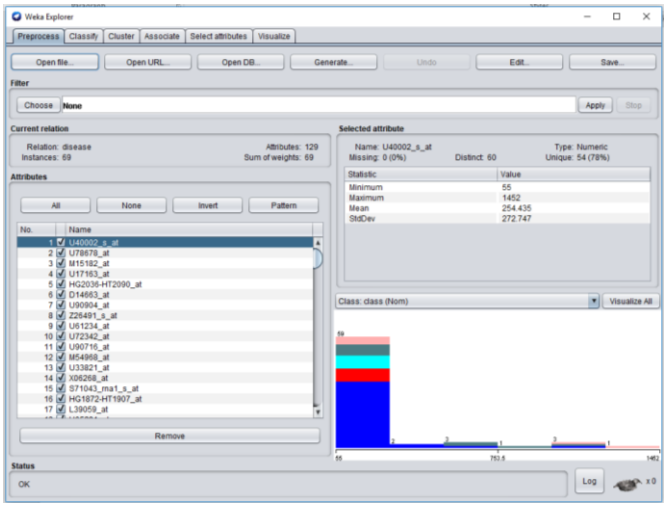
شماره دانشجویی : ۹۶۱۱۲۷۴

بهار ۹۸

در این پروژه قصد داریم از روی داده های میکرو آرایه های ژنتیک که برای آموزش در اختیار قرار گرفته و نوع بیماری نمونه های آن از قبل مشخص شده است استفاده شود تا نوع بیماری نمونه های دیگری از میکرو آرایه های ژنتیک که به عنوان فایل تست در اختیار قرار داده شده است با استفاده از یک کلاسیفایر منتخب پیش بینی گردد. تعداد نمونه های فایل آموزشی ۶۹ و تعداد نمونه های فایل تست ۲۳ مورد می باشد.

برای انجام این پروژه از دو برنامه استفاده شده است که یکی برای محاسبات، اصلاح و ساخت فایل های مورد نیاز در نرم افزار Weka که یک برنامه تحت ویندوز با زبان برنامه نویسی C# شخصا پیاده سازی شده است و دیگری خود نرم افزار Weka که برای بکارگیری الگوریتم های دسته بندی و دریافت خروجی مورد استفاده قرار گرفته اند. تصویر این دو نرم افزار را در زیر مشاهده می کنید:





کد برنامه تحت ویندوز نوشته شده در فولدر موجود SourceCode می باشد.

روند کارهای انجام شده در مراحل زیر توضیح داده شده است:

۱- پالایش داده ها

این مرحله توسط برنامه نوشته شده انجام می شود. در ابتدا سه فایل نمونه های آموزشی، فایل کلاس های نمونه های آموزشی و فایل نمونه های تستی به برنامه معرفی می شود. در ادامه با زدن دکمه خوتندن فایل ها برنامه داده ها را برای محاسبات پیش رو با فرمت خاص (جهت سهولت در انجام عملیات) در حافظه نگهداری می کند (متد buttonReadFiles_Click)

سپس با زدن دکمه Treshold در بخش پالایش داده ها ژن های نمونه آموزشی که از مقادیر ماکزیموم و مینیموم انتخابی (در اینجا مقادیر ۲۰ و ۱۶۰۰۰) حذف می شوند و دیگر در محاسبات مورد استفاده قرار نمی گیرند. (متد buttonTreshold_Click)

۲- انتخاب ژن های منتخب برای هر کلاس

این مرحله توسط برنامه نوشته شده انجام می شود. در این مرحله ابتدا بر اساس مقدار Fold Difference وارد شده (در اینجا مقدار ۲) در نمونه های آموزشی ژن ها حذف می شوند و در محاسبات بعدی مورد استفاده قرار نمی گیرند. سپس در هر کلاس برای ژن های باقی مانده مقدار T-Value محاسبه می شود و برای هر نمونه آموزشی تعداد ۲، ۴، ۶، ۸، ۱۰، ۱۲، ۱۵، ۲۰، ۲۵ و ۳۰ ژن با بیشترین T-Value انتخاب شده و با هم ترکیب می شوند(در محاسبات از مقدار قدر مطلق مقادیر T-Value استفاده شده است).
سپس برای هر تعداد ژن های منتخب فایل های مربوطه شامل فایل های زیر ساخته می شود.
(متد buttonCreateTopGenes_Click)

pp5i_train.top2.gr.csv
pp5i_train.top4.gr.csv
pp5i_train.top6.gr.csv
pp5i_train.top8.gr.csv
pp5i_train.top10.gr.csv
pp5i_train.top12.gr.csv
pp5i_train.top15.gr.csv
pp5i_train.top20.gr.csv
pp5i_train.top25.gr.csv
pp5i_train.top30.gr.csv

فرمت فایل های ژن های منتخب همانند فایل های آموزشی و تست با فرمت csv و به صورت ژن در سطر ساخته شده اند که برای استفاده شدن در برنامه Weka می بایست به فرمت arff با هدر فایل مخصوص شامل @Relation، @Attributes و @Data و به صورت ژن در ستون تبدیل شوند. که همزمان در برنامه این کار نیز انجام می شود و فایل های زیر نیز ساخته می شوند.

pp5i_train.top2.gr.arff
pp5i_train.top4.gr.arff
pp5i_train.top6.gr.arff
pp5i_train.top8.gr.arff
pp5i_train.top10.gr.arff
pp5i_train.top12.gr.arff
pp5i_train.top15.gr.arff
pp5i_train.top20.gr.arff
pp5i_train.top25.gr.arff
pp5i_train.top30.gr.arff

۳- تعیین بهترین کلاسیفایر و بهترین تعداد ژن منتخب

این مرحله در برنامه Weka انجام می شود. پس از ساختن فایل های با فرمت arff در مرحله قبل می توان این فایل ها را در برنامه Weka مورد استفاده قرار داد. برای این منظور به ازای متدهای درخواستی شامل NaiveBayes، J48، IB1، IBK=2,3,4 و همینطور یک الگوریتم به انتخاب شخصی که در اینجا RandomForest می باشد هر یک از فایل های ژن های منتخب را مورد ارزیابی قرار داده و میزان دقت و خطای هر یک را بدست می آوریم. نتایج این عملیات را در جدول زیر مشاهده می کنید:

RandomForest	IBK4	IBK3	IBK2	IB1	J48	NaiveBayes	
Correct : 69 Incorrect : 0 RMSE : 0.076	Correct : 65 Incorrect : 4 RMSE : 0.1394	Correct : 67 Incorrect : 2 RMSE : 0.1174	Correct : 66 Incorrect : 3 RMSE : 0.1211	Correct : 69 Incorrect : 0 RMSE : 0.027	Correct : 65 Incorrect : 4 RMSE : 0.1446	**Correct : 69 Incorrect : 0 RMSE : 0**	pp5i_train.top2
Correct : 69 Incorrect : 0 RMSE : 0.0743	Correct : 67 Incorrect : 2 RMSE : 0.1444	Correct : 65 Incorrect : 4 RMSE : 0.1346	Correct : 69 Incorrect : 0 RMSE : 0.1016	Correct : 69 Incorrect : 0 RMSE : 0.027	Correct : 67 Incorrect : 2 RMSE : 0.103	**Correct : 69 Incorrect : 0 RMSE : 0**	pp5i_train.top4
Correct : 69 Incorrect : 0 RMSE : 0.0753	Correct : 66 Incorrect : 3 RMSE : 0.1242	Correct : 67 Incorrect : 2 RMSE : 0.1198	Correct : 69 Incorrect : 0 RMSE : 0.0942	Correct : 69 Incorrect : 0 RMSE : 0.027	Correct : 67 Incorrect : 2 RMSE : 0.103	**Correct : 69 Incorrect : 0 RMSE : 0**	pp5i_train.top6
Correct : 69 Incorrect : 0 RMSE : 0.0719	Correct : 67 Incorrect : 2 RMSE : 0.1226	Correct : 67 Incorrect : 2 RMSE : 0.1085	Correct : 69 Incorrect : 0 RMSE : 0.0862	Correct : 69 Incorrect : 0 RMSE : 0.027	Correct : 67 Incorrect : 2 RMSE : 0.1031	**Correct : 69 Incorrect : 0 RMSE : 0**	pp5i_train.top8
Correct : 69 Incorrect : 0 RMSE : 0.0742	Correct : 66 Incorrect : 3 RMSE : 0.1255	Correct : 67 Incorrect : 2 RMSE : 0.1054	Correct : 69 Incorrect : 0 RMSE : 0.0942	Correct : 69 Incorrect : 0 RMSE : 0.027	Correct : 67 Incorrect : 2 RMSE : 0.1031	**Correct : 69 Incorrect : 0 RMSE : 0**	pp5i_train.top10
Correct : 69 Incorrect : 0 RMSE : 0.0704	Correct : 67 Incorrect : 2 RMSE : 0.1151	Correct : 67 Incorrect : 2 RMSE : 0.0989	Correct : 69 Incorrect : 0 RMSE : 0.0942	Correct : 69 Incorrect : 0 RMSE : 0.027	Correct : 68 Incorrect : 1 RMSE : 0.0726	**Correct : 69 Incorrect : 0 RMSE : 0**	pp5i_train.top12
Correct : 69 Incorrect : 0 RMSE : 0.0696	Correct : 67 Incorrect : 2 RMSE : 0.1118	Correct : 67 Incorrect : 2 RMSE : 0.0989	Correct : 69 Incorrect : 0 RMSE : 0.0862	Correct : 69 Incorrect : 0 RMSE : 0.027	Correct : 68 Incorrect : 1 RMSE : 0.0726	**Correct : 69 Incorrect : 0 RMSE : 0**	pp5i_train.top15
Correct : 69 Incorrect : 0 RMSE : 0.0702	Correct : 67 Incorrect : 2 RMSE : 0.1051	Correct : 67 Incorrect : 2 RMSE : 0.0989	Correct : 69 Incorrect : 0 RMSE : 0.0674	Correct : 69 Incorrect : 0 RMSE : 0.027	Correct : 68 Incorrect : 1 RMSE : 0.0726	**Correct : 69 Incorrect : 0 RMSE : 0**	pp5i_train.top20
Correct : 69 Incorrect : 0 RMSE : 0.0728	Correct : 66 Incorrect : 3 RMSE : 0.1179	Correct : 67 Incorrect : 2 RMSE : 0.092	Correct : 69 Incorrect : 0 RMSE : 0.0774	Correct : 69 Incorrect : 0 RMSE : 0.027	Correct : 68 Incorrect : 1 RMSE : 0.0726	**Correct : 69 Incorrect : 0 RMSE : 0**	pp5i_train.top25
Correct : 69 Incorrect : 0 RMSE : 0.071	Correct : 67 Incorrect : 2 RMSE : 0.1051	Correct : 67 Incorrect : 2 RMSE : 0.0956	Correct : 69 Incorrect : 0 RMSE : 0.0674	Correct : 69 Incorrect : 0 RMSE : 0.027	**Correct : 69 Incorrect : 0 RMSE : 0**	**Correct : 69 Incorrect : 0 RMSE : 0**	pp5i_train.top30

پس از بدست آوردن مقادیر دقت و خطای هر یک از الگوریتم ها روی فایل های ژن های منتخب شامل تعداد پیش بینی درست و غلط و همینطور مقدار خطای Root Mean Square Error مشخص شد که الگوریتم NiaveBayes برای کلیه فایل های منتخب و الگوریتم J48 برای فایل منتخب ۳۰ بهترین نتیجه را بدست می دهند. لذا به دلیل محاسبات کمتر فایل منتخب ۲ که در الگوریتم NaiveBayes نتیجه قابل قبولی را بدست داده است برای استفاده در پیش بینی نمونه های تستی انتخاب می شود.

۴- پیش بینی کلاس نمونه های تستی

در این مرحله از هر دو برنامه نوشته شده و برنامه Weka استفاده می شود. در ابتدا از برنامه نوشته شده در بخش تطبیق ژن های تست فایل ژن های متخب و همین طور تعداد آن را انتخاب نموده و دکمه ساخت فایل بهترین تست را می زنیم (متد buttonTestGeneExtrat_Click). در این عملیات بر اساس ژن های موجود در فایل منتخب، ژن های فایل نمونه های تستی نیز پالایش می شود یعنی تنها ژن هایی که در فایل منتخب موجود هستند باقی مانده و بقیه حذف می شوند که فایل

pp5i_test.best2.csv بدست می آید و در ادامه برای اینکه بتوان این اطلاعات را در برنامه Weka مورد استفاده قرار داد این فایل به فرمت arff تبدیل می شود و فایل pp5i_test.best2.arff بدست می آید.

در مرحله بعدی فایل ژن های منتخب به عنوان فایل پیش پردازش معرفی شده و فایل بدست آمده در بخش قبلی برای نمونه های تستی نیز به عنوان فایل تست در برنامه Weka جهت تعیین کلاس نمونه های تستی مورد استفاده قرار می گیرد. نتیجه این عملیات در فایل FarkhondehZeinali_Predicted.arff ذخیره شده است که نتیجه کلاس های پیش بینی شده برای نمونه های تستی را در جدول زیر مشاهده می کنید.

Class	Sample
MGL	101
EPD	102
MED	103
MED	104
EPD	105
MED	106
MED	107
MED	108
EPD	109
JPA	110
JPA	111
MED	112
MED	113
MED	114
MED	115
MED	116
MED	117
MED	118
MED	119
MED	120
RHB	121
MED	122
RHB	123