تمرین و پروژه شماره ۱ درس داده کاوی، خلاصه مقالات سؤال ۱

تکین جزایری، ۹۸۱۳۰۰۶ – امیرحسین رجبی، ۹۸۱۳۰۱۳

a) Student Academic Performance Prediction by using Decision Tree Algorithm

هدف کلی این مقاله ایجاد امکان پیشبینی نتایج تحصیلی دانشجویان است. برای این هدف، نویسندگان ۲۲ دانشجو از یک درس کارشناسی را تحت نظر گرفته اند و داده های مرتبط با آنها را ثبت کرده اند تا روی آن پردازش انجام دهند. همچنین این تحقیقات به کمک ابزار WEKA انجام گرفته است.

استخراج دانش از سامانه های مدیریت تحصیلی/درس (LMS/CMS) با نام داده کاوی آموزشی (EDM) شناخته می شود، با استفاده از متدهای گوناگونی قابل انجام است و در گذشته کارهای متعددی در زمینهٔ داده کاوی آموزشی انجام شده است که برخی از آنها هدف مشابه با این مقاله را دنبال کرده اند. این کارها از دو نظر با یکدیگر تفاوت دارند: متد مورد و ویژگی هایی از دانشجویان است که روی آنها بررسی صورت می گیرد.

در این مقاله، دو پارامتر در نظر گرفته شده است:

- اطلاعات آکادمیک دانشجویان: معدل کل، سابقهٔ افتادن در همان درس، سابقهٔ افتادن در تمامی دروس، مرکز موفقیت دانشجویی (SSC)، تکلیف ۱ و ۲، ارزیابی پایانی دوره، تعداد تقلب
- فعالیت های دانشجویی: میزان زمانی که دانشجو در سامانه Moodle می گذراند.

روش در مقاله به این صورت است که متدهای مختلف دسته بندی روی پارامترهای مورد بررسی اعمال می شود. از جملهٔ این روش های می توان به درخت تصمیم، جنگل تصادفی (Random Forest)، درخت مدل لجستیک (Logistic model tree)، بیز ساده و SMO اشاره کرد. در بررسی و مقایسهٔ روش های دسته بندی از دو فاکتور دقت و کاپا استفاده شده است. این الگوریتم ها توسط ابزار WEKA بر روی داده ها اعمال شده است ولی پیش از آن، کل مجموعهٔ داده ها (dataset) به صورت اسمی (nominal) تبدیل شده است.

نتیجهٔ این مقاله به اینصورت است که روش های جنگل تصادفی و همچنین SMO با داشتن دقت کامل توانسته اند با بیشترین دقت پیشبینی را انجام دهند. بدترین عملکرد در این مورد هم در الگوریتم های درخت REP، درخت مدل لجیستیک و درخت هوفدینگ (Hoeffding tree) با دقت ۴۵٪ رخ می دهد.

تحلیل شخصی از مقاله:

این مقاله دارای حجم نمونهٔ بسیار کم است که همین حجم نمونه هم بسیار متمرکز انتخاب شده اند (کل نمونه از یک کلاس گرفته شده است). به دلیل همین حجم پایین نمونه، مجموعهٔ اعتبارسنجی (validation set) تعریف نشده است و دقت از روی همان مجموعهٔ آموزش (training set) بدست آمده است.

همچنین از نظر روش کار با داده ها، به نظر می رسد اگر داده ها اسمی نشده بودند، الگوریتم درخت تصمیم می توانست عملکرد بهتری داشته باشد؛ زیرا در این حالت داده های پیوسته (مانند نمرات تکالیف یا نمرهٔ ارزیابی پایانی) را خود الگوریتم تقسیم بندی می کرد.

b) Classification Based on Decision Tree Algorithm for Machine Learning

این مقاله یک مقالهٔ مروری است که نگاهی اجمالی به درخت های تصمیم و برخی پیشرفت ها و تحولاتی که حول آنها اتفاق افتاده است، می پردازد. همچنین برخی مقالاتی را که کاربردی از درخت های تصمیم را بیان کرده است، بررسی می کند.

این مقاله ابتدا توضیحاتی کلی دربارهٔ یادگیری ماشین و دسته بندی (Classification) داده می شود. سپس درخت تصمیم به عنوان ابزاری برای دسته بندی معرفی شده و اطلاعاتی حول آن داده می شود. در ادامه مقداری دربارهٔ الگوریتم های مختلف ساخت درخت تصمیم، ایرادات و مزایای درخت تصمیم و همچنین معیارهای ارزیابی ناخالصی (Impurity) صحبت می شود.

ادامهٔ مقاله تماماً توضیح مختصر از مقالاتی است که در آنها از درخت تصمیم استفاده شده است و این مقالات از حیث حوزه مطالعات بسیار متنوع هستند که این تنوع می تواند مناسب بود الگوریتم درخت تصمیم را برای کاربردهای گوناگون نشان دهد. در بسیاری از این مقالات، از روش های دیگری مانند جنگل تصادفی، شبکه عصبی و KNN در کنار درخت تصمیم استفاده کرده اند. در ادامه به صورت تیتروار به بعضی از مقالات مورد بحث اشاره می کنیم:

- یکی از مقالات با استفاده از درخت تصمیم یک مطالعه روی بیماران دیابتی انجام داده است و در آخر روشی برای پیشبینی این بیماری ارائه می کند.
- مقالهٔ دیگر درخت تصمیم را برای تشخیص ارقام دست نوشته به کار می گیرد.
- یک مقاله برای استفاده از درخت تصمیم در تحلیل رفتارهای کاربران، ساختار جدیدی تحت عنوان درخت تصمیم رفتاری (BehavDT) معرفی می کند و با کار روی آن، به دقت ۹۰٪ می رسد.
- یکی از مقالات در تحلیل و کار روی سیگنال های ماهواره ای، از درخت های تصمیم کمک می گیرد.
- مقالهٔ دیگر یک نوع جدید از درخت تصمیم به نام XGBoost را طراحی می کند تا به کمک آن یک روش پیشبینی برای زمان معمول سیگارکشیدن ارائه دهد.
- در یکی از مقاله ها به تشخیص سلول های قرمز خونی معیوب توجه شده بود که این کار با یک درخت تصمیم با دقت بالایی انجام شده بود.