با یاد پروردگار

پروژه دوم درس شناسایی الگو – دانشکده کامپیوتر – دانشگاه علم و صنعت ایران

استاد درس: دکتر مرتضی آنالویی

تدریس یاران:

سید حسن طباطبائی – محمد بختیاری

Hassan.tbt1989@gmail.com educatemb@gmail.com

==========

طراحی رده‌بندهای مجمع برای داده‌های نامتعادل

مقدمه

امروزه در حوزه شناسایی الگو و یادگیری ماشین، مسائل با داده‌های نامتعادل(Imbalanced) اهمیت فراوانی دارند. طی سال‌های اخیر، متدها و رده‌بندهای مختلفی برای کار با این داده‌ها ساخته‌شده‌اند. از مهم‌ترین تکنیک‌های ابداع‌شده برای کار با این نوع داده‌ها و متعادل‌سازی چنین داده‌هایی، تکنیک SMOTE هست.

شرح

در این پروژه دانشجو باید چهار رده‌بند مختلف برای داده‌های نامتعادل را که مشخصات مقاله‌های آن در ادامه آورده می‌شود، مطالعه نموده و دقیقاً مطابق مقاله، به‌طور کامل و از پایه پیاده نموده و مجموعه داده اختصاصی خود را با آن‌ها تست کند و نتایج آن را به‌طور کامل شرح دهد. همچنین شیوه SMOTE که از آن در برخی از این رده‌بندها استفاده شده است نیز باید مطابق مقاله مربوطه پیاده شود.

این چهار رده‌بند عبارت‌اند از RB-Boost، SMOTE-Boost، RUS-BOOST، Ada-Boost M2

مشخصات مقاله ها

مشخصات مقاله‌ی هر روش که در بالا گفته‌شده است به شرح زیر هست:

Díez-Pastor, José F., et al. "Random Balance: Ensembles of variable priors classifiers for imbalanced data." *Knowledge-Based Systems* (2015).'

Chawla, Nitesh V., et al. "SMOTEBoost: Improving prediction of the minority class in boosting." *Knowledge Discovery in Databases: PKDD 2003*. Springer Berlin Heidelberg, 2003. 107-119.

Seiffert, Chris, et al. "RUSBoost: A hybrid approach to alleviating class imbalance." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40.1 (2009): 185-197.

Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." *icml*. Vol. 96. 1996.

Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

همچنین فایل PDF این مقاله‌ها در گروه درس، در اختیار شما قرار خواهد گرفت.

موارد خواسته‌شده

قسمت اول – اجباری – ۱۰۰ امتیاز

به هرکدام از دانشجویان یک مجموعه داده اختصاص داده‌شده است که پس از پیاده‌سازی این چهار الگوریتم، باید آن را بر روی این مجموعه داده بیازماید و آن را با دو رده‌بند دیگر، SVM و Random Forest مقایسه کند. این آزمون باید به‌صورت 5-Fold صورت بگیرد(۴ قسمت یادگیری و ۱ قسمت تست) و دقت تست میانگین هر۵ تکرار، گزارش شود. این مقایسه‌ها باید برای سه اندازه مجمع ۱۰، ۵۰ و ۱۰۰ تست شود. و در جدول آورده شده و نمودار آن رسم شود. برای رده‌بند پایه از یک درخت پیاده‌سازی شده موردنظر خود که قابلیت محاسبه احتمال را داشته باشد، می‌توانید استفاده کنید (مانند C4.5 ، Cart یا Decision Stump) در صورت استفاده از هرکدام از این درخت‌ها یا هر رده‌بند پایه دیگر، دلیل خود را برای استفاده از آن، بیان کنید. معیارهای Precision، Recall را برای این۶ رده‌بند به تفکیک اندازه مجمع در نمودار میله ای رسم کرده و باهم مقایسه کنید(هر معیار و اندازه در یک نمودار) . همچنین برای نمودار ROC نیز همین کار را انجام دهید. درنهایت با توجه به نتایج به‌دست‌آمده تحلیل خود را از کارایی هرکدام از رده‌بندها ارائه دهید.

قسمت دوم – اختیاری – ۴۰ امتیاز

در این قسمت برای مقایسه باید از روش‌های آماری استفاده شود. به این شکل که ۱۰ بار به‌صورت تصادفی، داده‌ها به نسبت ۷۰ به ۳۰ درصد به Train و Test تقسیم‌شده و رده‌بندها را روی آن اجرا نمایید. برای هر رده‌بند، ۱۰ دقت به دست می‌آید. سپس با استفاده از آزمون ANOVA، مشخص شود که آیا یکی از روش ها به‌طور معناداری بهتر است؟ اگر بله کدام‌یک؟ توجه کنید که باید به‌دقت، آزمون را پیاده‌سازی نموده، اجرا کرده و همچنین نتایج و شیوه اجرا را باید به‌طور کامل و مرحله‌به‌مرحله در گزارش توضیح دهید.

می‌توانید از آموزش موجود در لینک زیر بهره ببرید:

https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/

شیوه پیاده‌سازی و موارد تحویلی

برای پیاده‌سازی این پروژه می‌توانید از زبان برنامه‌نویسی دلخواه خود استفاده کنید. همچنین می‌توانید برای رده‌بندهای پایه، و دو رده‌بندی که مقایسه با آن‌ها انجام می‌گیرد(SVM و Random Forest)، از کتابخانه‌های پیش‌ساخته آن زبان بهره ببرید. در گزارش خود، ابتدا مختصری از دیتاست، مختصری درباره SMOTE و RBBOOST، شیوه پیاده‌سازی و نتایج و نمودارها را بیاورید. همچنین، کد نرم‌افزاری کامل نیز به همراه گزارش در یک پوشه به فرمت

IUSTPR982-StudentFullName-StudentNumber

قرارگرفته و به‌صورت زیپ شده به ایمیل hassan.tbt1989@gmail.com حداکثر تا پایان مهلت پروژه دوم(۱۵ دی‌ماه) ارسال شود. همچنین در روز تحویل(که تاریخ آن متعاقباً اعلام می‌شود) گزارش و کدهای ارسالی، توسط دانشجو کاملاً شرح داده‌شده و اجرا می‌شود.

** دانشجویان عزیز سؤالات خود را فقط از طریق گروه تلگرامی بپرسند تا سایر دوستان نیز در صورت نیاز از پاسخ‌ها بهره ببرند.

مجموعه داده اختصاصی هر دانشجو

http://sci2s.ugr.es/keel/imbalanced.php

این قسمت تکمیل خواهد شد...