مقایسه روش‌های Bagging و Boosting برای دسته‌بندی داده‌ها

مهدی نصیری[۱]، علی هادیان[۲]، ، بهروز مینایی[۳]

چکیده

در این مقاله دقت روش‌های Bagging و Boosting برای دسته‌بندی داده‌ها با توجه به معیار "دقت" مقایسه شده است. تاثیر عواملی مانند اندازه مجموعه داده، تعداد صفات دودویی و عددی و درصد داده استفاده شده برای آزمون در شرایط استقلال مجموعه داده از مسئله خاص، مورد بررسی قرار گرفته است.

نتایج بدست آمده بیانگر آن است که در اکثر موارد بهتر از روش دیگر عمل می‌نماید. برای یک داده خاص روش‌ها در دو حالت بررسی شده است. یک حالت زمانی است که داده آموزشی و داده آزمایشی یکی باشند و حالت دیگر ۳۰٪ از داده آزمایشی و ۷۰٪ آموزشی باشد.

کلمات کلیدی

Bagging و Boosting ، مجموعه داده‌ها، داده‌های آموزشی، داده‌های آزمایشی

۱. مقدمه

اگر اندازه داده آموزش در مقایسه با بعد داده‌ها کوچک باشد، داده آموزش اغلب ممکن است توزیع واقعی داده را نمایش ندهد. لذا مدل طبقه‌بندی که بر طبق این داده آموزشی ساخته می‌شود ، ممکن است تحت تاثیر قرار گرفته و اختلاف[۱] بزرگی پیدا کند. در نتیجه این طبقه‌بندی‌کننده ضعیف کارایی ضعیفی خواهد داشت. به منظور بهتر کردن طبقه‌بندی‌کننده ضعیف توسط تثبیت کردن تصمیمش ، روش‌هایی از قبیل تنظیم[۲] یا تزریق اختلال[۳] استفاده می‌شود. روش دیگر برای بهبود طبقه‌بندی‌کننده ضعیف، استفاده از ترکیب طبقه‌بندی‌کننده‌هایی که از نسخه‌های تعدیل شده[۴] داده آموزشی اصلی[۵] بدست آمده (مثلا توسط نمونه گیری[۶] یا وزن دهی[۷]) ، می‌باشد. این روش در روش باد کردن[۸] و ترقی دادن[۹] پیاده‌سازی می‌شود. در روش باد کردن، هر شخصی از داده آموزشی نمونه‌برداری می‌کند، خودراه‌انداز[۱۰] مستقل تصادفی مولد داده آموزشی را کپی می‌کند[۱۱]، بر روی هر یک از این داده آموزشی که توسط خودراه انداز بوجود آمده یک طبقه‌بندی‌کننده می‌سازد و آن‌ها را توسط حداکثر آرای[۱۲] ساده در قانون تصمیم نهایی جمع می‌کند[۱۳]. در روش بادکردن، طبقه‌بندی‌کننده‌ها بر روی نسخه‌های وزن‌دهی شده داده آموزشی ساخته می‌شوند، که به طور پی در پی در الگوریتم بدست آورده می‌شوند. در ابتدا ، تمام اشیاء به وزن‌های مساوی دارند، و اولین طبقه‌بندی‌کننده بر اساس این مجموعه داده ساخته می‌شود. سپس وزن‌ها بر طبق کارایی طبقه‌بندی‌کننده تغییر داده می‌شوند. شی‌های طبقه‌بندی‌شده نادرست وزن‌های بیشتری می‌گیرند و طبقه‌بندی‌کننده بعدی بر روی داده آموزشی مجدد وزن‌دهی شده اعمال می‌شود. در این روش یک ترتیبی از مجموعه‌های آموزش و طبقه‌بندی‌کننده‌ها بدست آورده می‌شود، که سپس توسط حداکثر رای‌گیری ساده یا وزن‌دهی شده در تصمیم نهایی ترکیب می‌شوند.

۲. بگینگ

بگینگ (متراکم شدن خودکار) توسط لئو بریمن در سال ۱۹۹۴ پیشنهاد شد که برای بهبود دادن رده بندی توسط ترکیب کردن رده بندی های مجموعه‌های آموزشی به طور تصادفی تولید شده، می‌باشد.

این روش یک یک متا الگوریتم می‌باشد که برای بهبود دادن یادگیری ماشین رده‌بندی و مدل‌های پسرفتی بر حسب پایداری و دقت رده‌بندی می-باشد. این روش همچنین واریانس را کاهش داده و به دوری از اورفیتینگ کمک می‌کند. اگر چه این روش معمولا در دخت تصمیم به کار می رود اما می تواند در هر نوع مدل استفاده شود. بگینگ یک حالت مخصوص از روند مدل میانگین می‌باشد.

[۱] دانشجوی کارشناسی ارشد هوش مصنوعی دانشگاه علم و صنعت ایران nasiri@comp.iust.ac.ir
[۲] دانشجوی کارشناسی مهندسی نرم‌افزار دانشگاه علم و صنعت ایران hadian@comp.iust.ac.ir
[۳] استادیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران b_minaei@iust.ac.ir

یک مجموعه آموزشی استاندارد D به اندازه n را فرض کنید، بگینگ توسط نمونه گیری به طور یکنواخت و با جایگزینی مثال‌ها از D، m مجموعه آموزشی جدید D_i با اندازه $n \geq n'$ تولید می شود. نمونه گیری با جایگزینی این امکان را می دهد که بعضی از مثال ها امکان تکرار در هر D_i را داشته باشند. اگر $n = n'$ باشد لذا برای n بزرگ ، مجموعه D_i انتظار داشتن ٪۶۳.۲ از مثال‌های بیهمتای D را دارد و بقیه مثال ها تکراری می باشند. این نوع نمونه‌گیری به عنوان نمونه گیری خورراه‌انداز شناخته می‌شود. m مدل برای استفاده کردن m نمونه‌های خودکار بالا گنجانیده شده و این مدل‌ها توسط متوسط گیری خروجی (برای پسرفت) یا رای گیری (برای رده بندی) ترکیب می‌شوند.

از آنجاییکه این روش چندین پیشگویی کننده را میانگین می‌گیرد، لذا برای بهبود مثال‌های خطی مفید نمی‌باشد.

۳. بوستینگ

بوستینگ یک متا الگوریتم یادگیری ماشین برای اجرای یادگیری نظارت شده می باشد. بوستینگ بر سوال مطرح شده توسط کیرنز بنا شده است: آیا یک مجموعه یادگیرنده‌های ضعیف می‌تواند یک یادگیرنده واحد قوی بسازد؟ یک یادگیرنده ضعیف یک رده بندی کننده‌ای تعریف می‌شود که فقط اندکی با رده بندی صحیح همبسته است. در حقیقت، یک یادگیرنده قوی یادگیرنده‌ای است که به طور دلخواهانه همبسته‌ی خوبی با رده بندی صحیح دارند.

پاسخ مثبت به سوال کیرنز یک انشعاب‌های مهم در یادگیری ماشین و آمار دارد.

۳.۱. الگوریتم‌های بوستینگ

تا موقعی که بوستینگ به صورت الگوریتمی تحمیل نشود، اکثر الگوریتم‌های بوستینگ عبارتند از به طور تکراری یاد گرفتن رده بندی کننده‌های ضعیف نسبت به توزیع و اضافه کردن آن‌ها به رده‌بندی کننده قوی نهایی. موقعی که آن‌ها اضافه می‌شوند، نوعا در بعضی روش‌هایی وزن‌دهی می‌شوند که معمولا با دقت یادگیرنده ضعیف مرتبط است. بعد از اضافه کردن یک یادگیرنده ضعیف، داده دوباره وزن دهی می شود: مثال‌هایی که اشتباه رده-بندی شوند وزن بیشتری بدست آورده و مثال‌هایی که به درستی رده‌بندی شوند وزن از دست می‌دهند (بعضی الگوریتم‌های بوستینگ عملا وزن مثال-های مکررا نادرست رده بندی شده را کاهش می دهند مانند بوست توسط اکثریت و بوست خرمایی). بنابراین، یادگیرنده‌های ضعیف آینده بیشتر بر مثال‌هایی تمرکز می‌کند که یادگیرنده‌های ضعیف قبلی به نادرستی رده بندی کردند.

تعداد الگوریتم‌های بوستینگ زیادی وجود دارد. الگوریتم‌های اصیل، پیشنهاد شده توسط رابر اسچاپیر (فرموله کردن درجه اکثریت بازگشتی) و یوآو فروند (بوست توسط اکثریت)، انطباق پذیر نبودند و نتوانستند فایده‌ی کاملی از یادگیرنده‌های ضعیف بگیرند.

فقط الگوریتم‌هایی که در قاعده یادگیری محتملا تقریبا صحیح الگوریتم‌های بوستینگ قابل اثبات هستند، الگوریتم‌های بوستینگ می‌باشند. الگوریتم‌های دیگر که در روح با الگوریتم‌های بوستینگ شبیه هستند گاهی اوقات "الگوریتم‌های اهرمی" نامیده می‌شوند، هرچند آن‌ها گاهی اوقات نادرست الگوریتم‌های بوستینگ صدا زده می‌شوند.

۳.۲. آدابوست

آدابوست، مختصر شده از بوستینگ انطباقی، یک الگوریتم یاد ماشین هست، توسط یوآو فروند و رابرت اسچاپیر به شکل قاعده درآورده شد. آن یک متا الگوریتم می‌باشد و می تواند در ترکیب با تعداد زیادی الگوریتم‌های یادگیری برای بهبود کارایی‌شان استفاده شود. آدابوست تا حدی وقف پذیر است که ساخت رده‌بندی‌کننده‌های بعدی برای آن نمونه‌هایی که توسط رده‌بندی‌کننده‌های قبلی نادرست رده‌بندی شدند تنظیم شود. آدابوست به داده‌های نویزدار و بخش مجزا حساس می‌باشد. در غیر اینصورت، آن در مسائل اورفیتینگ حساسیت کمتری نسبت به الگوریتم‌های یادگیری دیگر دارد.

آدابوست مکررا در سری‌های گرد کردن $t = ۱, \ldots, T$ یک رده‌بندی‌کننده ضعیف نامیده می‌شود. برای هر فراخوانی یک توزیع وزن‌های D_t بروز رسانی می‌شود که اهمیت مثال‌ها را برای رده‌بندی در مجموعه داده مشخص می‌کند. در هر گرد کردن، وزن‌های هر مثالی که به نادرستی رده بندی شده افزایش می‌یابد (یا به طور جایگزین، وزن‌های هر مثالی که به درستی رده‌بندی شده کاهش می‌یابد)، به‌طوری‌که رده‌بندی‌کننده جدید بیشتر بر روی این مثال‌ها رده‌بندی می‌کند.

۴. نتیجه گیری

Bagging مفهومی برای ترکیب رده بندی های پیش بینی شده از چند مدل به کار می رود. فرض کنید که قصد دارید مدلی برای رده بندی پیش بینی بسازید و مجموعه داده های مورد نظرتان کوچک است. شمامی توانید نمونه هایی(با جایگزینی) را از مجموعه داده ها انتخاب و برای نمونه های حاصل ازدرخت رده بندی استفاده نمایید. به طورکلی برای نمونه های مختلف به درخت های متفاوتی خواهید رسید. سپس برای پیش بینی با کمک درخت های متفاوت به دست آمده از نمونه ها ،یک رای گیری ساده انجام دهید.رده بندی نهایی ، رده بندی ای‌خواهد بود که درخت های مختلف آنرا پیش بینی کرده‌اند.

Boosting مفهومی برای تولید مدل‌های چندگانه (برای پیش بینی یا رده بندی) به کار می‌رود. Boosting نیز از روش C&RT یا CHAID استفاده و ترتیبی از رده‌بندها را تولید خواهد کرد.

در جدول ۱ مشخصات مجموعه داده‌ها با توجه به نوع صفات مشخص شده است. در جدول ۲ نتایج دو روش بگینگ و آدابوستینگ با توجه به ۳ مجموعه داده آموزشی مشخص شده است. در این جدول مجموعه داده‌ها آموزشی و آزمایشی یکی هستند. در جدول شماره ۳ از ۷۰٪ مجموعه داده‌های جدول ۱ بعنوان داده‌های آموزشی و ۳۰٪ بعنوان داده آزمایشی انتخاب شده است. با توجه به نتایج می‌توان گفت:

- اعتماد به دقت روش بوستیگ در داده‌هایی که صفات آن فقط عددی است بیشتر از بگینگ است. (با توجه به اینکه داده آزمایشی جز داده آموزشی باشد یا نه)
- اعتماد به دقت داده‌هایی که نوع صفات آن اسمی است در بگینگ بیشتر از بوستینگ است.(با توجه به اینکه داده آزمایشی جز داده آموزشی باشد یا نه)
- دقت داده‌هایی که نوع صفت اسمی در آنها وجود دارد در بگینگ بیشتر از بوستینگ است. این در حالی است که اگر داده‌های آزمایشی بخشی از داده‌های آموزشی باشد دقت بوستینگ بیشتر است.
- دقت داده‌هایی که نوع صفات آن فقط عددی است در بوستینگ بیشتر از بگینگ است.
- دقت داده‌هایی که نوع صفات آن فقط اسمی است در هر دو روش برابر است.
- در داده‌هایی که هر دو نوع صفت اسمی و عددی وجود دارد نیز دقت‌ها تقریبا برابر است و دقت بگینگ اندکی بیشتر از بوستینگ است.

جدول ۱- مشخصات مجموعه دادها

مجموعه داده	تعداد صفت اسمی[۱۵]	تعداد صفت عددی[۱۴]	نوع رده‌بند
(iris)۱	۰	۴	اسمی
(contact-lenses)۲	۲	۲	اسمی
(laboar)۳	۸	۸	اسمی
(weather)۴	۴	۰	اسمی

جدول ۲- مقایسه با داده های آموزشی و آزمایشی یکسان

مجموعه داده	Bagging	Adaboosting
۱	۹۷.۳۳	۹۵.۳۳
۲	۶۴.۲۸	۱۰۰
۳	۹۴.۹۳	۹۸.۲۵
۴	۶۴.۲۸	۸۵.۷۱

جدول ۳- مقایسه در حالت:۷۰٪ داده ها به صورت آموزشی و ۳۰٪ به عنوان داده آزمایشی

مجموعه داده	Bagging	Adaboosting
۱	۷۸.۵	۹۲.۶
۲	۷۴.۷۳	۷۳.۷۶
۳	۶۵.۲۸	۶۱.۵۳
۴	۶۴	۶۴

۵. مراجع

[۱] Marina Skurichina and Robert P.W.Duin; "*The Role of Combining Rules in Bagging and Boosting*", ۲۰۰٤.

[۲] Ayhan Demiriz, Kristin P. Bennett, and John Shawe-Taylor; "*Linear programming boosting via column generation*", Machine Learning, ٤٦(۱/۲/۳):۲۲٥–۲٥٤, ۲۰۰۲.

[۳] Thomas G. Dietterich; "*An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization*", Machine Learning, ٤۰(۲):۱۳۹–۱٥۸, ۲۰۰۰.

[۴] Harris Drucker; "*Improving regressors using boosting techniques*", In Machine Learning: Proceedings of the Fourteenth International Conference, pages ۱۰۷–۱۱٥, ۱۹۹۷.

[۵] Yoav Freund; "*Boosting a weak learning algorithm by majority*", Information and Computation, ۱۲۱(۲):۲٥٦–۲۸٥, ۱۹۹٥.

[۶] Yoav Freund. An adaptive version of the boost by majority algorithm. Machine Learning, ٤۳(۳):۲۹۳–۳۱۸, June ۲۰۰۱.

[۷] J. Han, M. Kamber; "*Data Mining: Concepts and Techniques*", Second edition, Elsevier, ۲۰۰٦.

[۸] Pang-Ning Tan, Michael Steinbach, Vipin Kumar; "*Introduction to Data Mining*", Addison-Wesley; ۲۰۰٦.

[۹] Adam Fadlalla; "*An experimental investigation of the impact of aggregation on the performance of data mining with logistic regression*", Elsevier, Information & Management, pp. ٦۹٥–۷۰۷, ۲۰۰٥.

[۱۰] J. Huang, J. Lu,C.X. Ling; "*Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy*"; Proceedings of the Third IEEE International Conference on Data Mining , ۲۰۰۳.

[۱۱] L. Xu, M-Y. Chow, X. Z. Gao; "*Comparisons of Logistic Regression and Artificial Neural Network on Power Distribution Systems Fault Cause Identification*"; IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications, Finland, June ۲۸-۳۰, ۲۰۰٥.

[۱۲] N.B. Amor, S. Benferhat, Z. Elouedi; "*Naive Bayes vs decision trees in intrusion detection systems*", Proceedings of the ۲۰۰٤ ACM Symposium on Applied Computing, Nicosia, Cyprus ۲۰۰٤.

[۱۳] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura; "*A comparative study of K-Nearest Neighbour, Support Vector Machine and Multi-Layer Perceptron for Thalassemia screening*", Chemometrics and Intelligent Laboratory Systems, pp. ۱۳– ۲۰, ۲۰۰۳.

[۱۴] Yong Soo Kim; "*Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size*", Elsevier, Expert Systems with Applications, pp.۱۲۲۷–۱۲۳٤, ۲۰۰۸.

[۱٥] C. J. C. Burger; "*A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery*", Bell Labs/Lucent, ۲(۲):۹٥٥-۹۷٤, ۱۹۹۸.

[۱۶] Haykin; "*Neural Networks: A Comprehensive Foundation*", Second Edition, Prentice-Hall Inc., ۱۹۹۹.

[۱۷] V. N. Vapnick, "*The Nature of Statistical Learning Theory*", Second Edition, Springer-Verlag New York Inc.,۲۰۰۰.

[۱۸] Ian H. Witten, Eibe Frank; "*Data Mining: Practical Machine Learning Tools and Techniques*", Second Edition; Elsevier, Morgan Kaufmann publications; ۲۰۰٥.

---

۱ Variance

۲ Regularization

۳ Noise Injection

۴ Modified Versions

۵ Original Training Set

۶ Sampling

۷ Weighting

۸ Bagging

٩ Boosting

١٠ Bootstrap

١١ Replicate

١٢ Majority Vote

١٣ Aggregate

١٤ numeric

١٥ nominal