# روش یادگیری نظارتی برای مجموعه داده های نامتوازن

# چکیده

این مقاله روش یادگیری جدیدی برای برنامه های دسته بندی نمونه که با مجموعه داده های نامتوازن درگیرند ارائه می کند. در این روش، یک تکنیک خوشه بندی به کار گرفته می شود تا مجموعه اصلی آموزشی را در یک مجموعه کوچکتر که شامل نماینده نسخه آموزشی است، و توسط مراکز خوشه بندی وزین و خروجی های هدف آنها ارائه شده، دوباره نمونه گیری کند. بر اساس روش یادگیری پیشنهاد شده، چهار الگوریتم آموزشی برای شبکه های عصبی پیش خور استنتاج شده است. این الگوریتم ها در سه مجموعه داده ها ارزیابی مقایسه ای و آزمایش شده اند. نتایج آزمایشی نشان می دهند که با روش یادگیری ارائه شده، امکان دارد شبکه هایی طراحی کرد تا بر مشکلات دسته نامتوازن فائق شد، بدون اینکه با کارکرد تمام دسته بندی به مصالحه پرداخت.

# 1. مقدمه

با آنکه پیشرفت قابل ملاحظه ای در دسته بندی الگوها صورت گرفته است، هنوز موارد بسیاری باقی مانده است. مسئله ای که در این مقاله روی آن تمرکز خواهیم کرد این است که چطور کار دسته بندی را از مجموعه داده های نامتوازن بیاموزیم. مسئله دسته نامتوازن، که یکی از مسائل اساسی در یادگیری ماشین است، اخیرا توجه زیادی به خود معطوف کرده است. در بسیاری از کاربردهای خطایابی دنیای واقعی، برای مثال امنیت کامپیوتری، بیوپزشکی، و مهندسی، توزیع ناهموار الگوهای داده ها بسیار رایج است، که در آنها تعدادی از موارد آموزشی در یک دسته اقلیت در مقایسه با دسته های اکثریت دیگر بسیار کوچکتر است؛ در نتیجه، دسته بندی کننده متمایل خواهد بود به دسته اکثریت توجه کند. یک تحقیق که توسط مورفی و همکاران صورت گرفته نشان داد که شبکه های عصبی سنتی که پیش خور می شوند در یادگیری از مجموعه داده های نامتوازن مشکل دارد. به دلیل موارد بیش از اندازه قوی آموزشی، شبکه تمایل پیدا می کند دسته های اقلیت را نادیده بگیرد و با آن مانند یک اختلال رفتار کند. به طور کلی، یادگیری الگوریتمها برای مشکلات نامتوازن دسته می تواند به دو گروه تقسیم شود: نمونه گیری دوباره و بر اساس هزینه حساس. روشهای نمونه گیری دوباره مانند نمونه گیری بیش از حد و نمونه گیری بسیار کم احتمال اولیه دسته اکثریت و اقلیت در یک مجموعه آموزشی را برای بدست آوردن یک تعداد موارد بیشتر توازن یافته در هر دسته را تعدیل می کند. روش نمونه گیری بسیار کم یک مجموعه کوچکتر موارد اکثریت را استخراج می کند درحالیکه تمام موارد اقلیت را حفظ می کند. این روش برای کاربردهایی با مقیاس بزرگ مناسب است که در آن تعداد نمونه های اکثریت عظیم است و کم کردن موارد آموزشی زمان آموزش را تقلیل می دهد و مشکلات یادگیری را بیشتر مهار می کند. با این وجود، یک مسئله که به همراه تکنیک نمونه گیری بسیار کم وجود دارد این است که ممکن است مواردی از اطلاعات را از موارد دورانداخته شده از دست بدهیم.

در مقایسه با نمونه گیری بسیار کم، روش نمونه گیری بیش ازحد تعداد موارد اقلیت را با نمونه گیری بیش از حد از آنان ازدیاد می بخشد. سود این روش این است که هیچ اطلاعاتی از نمونه های آموزشی از دست نمی رود زیرا تمام موارد به کار گرفته می شوند. با این وجود، موارد اقلیت بیش از اندازه در مجموعه آموزشی مطرح می شوند و به علاوه، اضافه کردن موارد آموزشی به معنی ازدیاد زمان آموزشی است. روشهای بر پایه موازنه نوع دیگری از تکنیک نمونه گیری بیش از حد است، که در آن وزنهای بیشتری برای موارد آموزشی اقلیت واگذار می شود. روشهای برپایه حساسیت هزینه عملیات جایگزین مسئله عدم توازن دسته است. برادی و زانگ شبکه های عصبی حساس به هزینه را بوسیله واگذاری هزینه های متفاوتی به خطاها در دسته های مختلف معرفی می کنند. روش آنها صحت دسته بندی برای دسته های اقلیت را بوسیله واگذاری هزینه بیشتری به آنها ارتقا می بخشد. با این وجود، اضافه کردن یک تابع هزینه در مرحله یادگیری امکان توزیع را تعدیل می کند.

در این مقاله، روش جدیدی برای یادگیری پیش بینی شده با مجموعه داده های نامتوازن را معرفی می کنیم. مفهوم روش ما شبیه به نمونه گیری بسیار کم است. با این وجود، دسته بندی نظارت نشده را برای کم کردن موارد آموزشی اکثریت، بوسیله انتخاب مراکز خوشه بندی به عنوان نماینده نمونه ها به کار می گیریم. به علاوه، موازنه نسخه های آموزشی اقلیت و اکثریت در تابع هزینه معرفی شده است؛ وزن هر دسته به عنوان تقریب انبوه احتمال آن به کار گرفته می شود. بقیه مقاله به این صورت سازمان یافته است. بخش 2 روش یادگیری پیشنهادی را توضیح می دهد و چهار الگوریتم یادگیری بهبود یافته بر مبنای نزول شیب، نزول شیب با اندازه حرکت، پس-انتشار انعطاف پذیر و لونبرگ مارکارد را استنتاج می کند. بخش 3 نتایج آزمایش را ارائه می کند که در آن روش پیشنهاد شده برای دسته بندی های مختلف به کار برده می شود. سرانجام، بخش 4 نتایج را ارائه می دهد.

# 2. الگوریتمهای آموزشی بهبود یافته برای شبکه های عصبی پیش خور

فرض کنید یک شبکه عصبی پیش خور چند لایه باید با استفاده از یک مجموعه M چندتایی {xm, dm}، که در آن m= 1,2,…,M; xm بردار ورودی و dm بردار خروجی متناظر یا بردار مطلوب است، آموزش ببیند. بردار w را برداری در نظر می گیریم که شامل تمام پارامترهای آزاد شبکه است، شامل وزنها و تمایل ها. هدف یادگیری نظارت شده این است که بردار w0 را که تابع هزینه را کمینه می کند پیدا کند. یک تابع هدف *میانگین مربعات خطا* است (MSE)، که اینگونه تعریف می شود



که در آن N تعداد نرون های لایه خروجی است، و yim خروجی شبکه است. زمانی که تعداد موارد آموزشی دسته های مختلف ناهموار می شود، سهم خطای هر دسته برای تابع هدف نامساوی است. در مسئله دو- دسته ای، دسته اکثریت اثر بسزایی بر جریان بهینه سازی دارد. بنابراین، ما برای شبکه های عصبی پیش خور الگوریتمی که بیشتر تاثیرگذار باشد پیشنهاد می کنیم. در این روش، یک مرحله پیش پردازی معرفی شده است تا تعداد متعادل تری از نمونه ها در هر دسته کسب شود. در این مرحله پایانی، خوشه سازی نظارت نشده برای نمونه های یادگیری دسته های اکثریت به کار گرفته می شود تا مراکز خوشه که نمایندگی فشدسته دسته های اکثریت را حاصل می کنند، استخراج کند.

در اینجا، خوشه سازی به طور مستقل به تمام نمونه های آموزشی که نماینده یک دسته خاص هستند اعمال می شود. بنابراین، هر خوشه، نمونه هایی را از یک دسته نشان می دهد، و هر دسته بوسیله چند خوشه نمایان می شود. در این روش، ما با مجموعه اطلاعات نامتوازن کار خواهیم کرد، به این وسیله که به راحتی همان تعداد خوشه را برای هر دسته واگذار می کنیم. بعد از خوشه سازی، مجموعه اطلاعات به نمونه k تقلیل می یابد، که هر کدام بوسیله یک خوشه *مرکز ثقل ck و سایز* مشخص می شوند. در اینجا، خوشه سایز zk به راحتی تعداد نمونه های آموزشی در خوشه است. در ادامه، چهار الگوریتم آموزشی که سایز خوشه و مراکز ثقل در قانون آموزشی را مجتمع می کنند ارائه می دهیم.

## 2.1 تابع هزینه بهبود یافته

در مرحله یادگیری نظارت شده، نمونه های آموزشی بوسیله یک مجموعه مراکز ثقل خوشه که سپس به شبکه به همراه خروجی های هدف ارائه می شوند، جایگزین می شوند. برای جبران اطلاعاتی در طول جریان خوشه بندی از دست رفته اند، وزنهای هر کلاس در تابع هزینه معرفی می شوند:



که در آن dik i-امین عنصر هدف یا بردار خروجی dk است، و pk وزن خوشه ای است که جرم احتمال را به طور تقریبی بیان می کند،



که در آن Ncl تعداد دسته های مجموعه آموزشی است، ωi سایز دسته i است، و γki درجه عضویت خوشه k در دسته i است:



## 2.2 الگوریتمهای آموزشی بهبود یافته

الگوریتمهای بهینه سازی بسیاری برای کم کردن E می توانند برای تعلیم شبکه های عصبی پیش خور استنتاج شوند. در این مقاله، ما چهار الگوریتم را اجرا می کنیم، به نامهای نزول شیب (GD)، نزول شیب با اندازه حرکت و درجه یادگیری متغیر (GDMV)، پس-انتشار انعطاف پذیر (RPROP)، و لونبرگ مارکارد (LM), که بر پایه هدف روش ما هستند. هر الگوریتم آموزشی اوزان شبکه ها و تمایلاتشان را بر اساس w(t+1)=w(t)+∆w(t) به روز رسانی می کند. زیرا جزئیات الگوریتمهای استاندارد شده در {13، 6، 12، 5} یافت می شوند، ما اینجا فقط خصوصیات اصلی آنها را خلاصه می کنیم.

* نزول شیب: وزنها در امتداد شیب منفی به روزرسانی می شوند ، که در آن 𝛼 نرخ یادگیری اسکالر است.
* GD با اندازه حرکت و نرخ یادگیری متغیر: به روز رسانی وزن در امتداد یک شیب منفی انجام و به روز رسانی قبلی وزن است



که در آن 𝜆 پارامتر مقدار حرکت است،1>λ>0، و 𝛼(t) نرخ یادگیری اسکالر سازگار است.

* پس انتشار انعطاف پذیر: به روز رسانی وزن فقط به علامت شیب بستگی دارد



که در آن i ­(t)Δ گام سازگار مخصوص به وزن ωi است.

* لونبرگ مارکارد: قانون به روز رسانی وزن به این صورت ارائه می شود



که در آن ρ ماتریس وزنی خوشه بسط یافته است. برای جزئیات درمورد ρ به {11} مراجعه کنید.



جدول 1. مقایسه الگوریتمهای استاندارد و بهبود یافته بر روی مجموعه داده های مبنا

# 3. آزمایشات و تحلیلها

در این بخش، روش یادگیری هدفدار را برای سه مسئله محک به کار می گیریم، که از منبع اطلاعات UCI گرفته شده است. محک ها مجموعه داده های نارسایی کبدی، هپاتیت و دیابتهای سرخپوستان بودند. جزییات این مجموعه داده ها در جدول 2 خلاصه شده اند. هدف ما مطالعه تعمیم قابلیتهای روش پیشنهاد شده است، در مقایسه با روش استاندارد آموزش شبکه های عصبی. مقایسه بر اساس وارسی اعتبار پنج جانبه در کار دسته بندی است. برای هر جنبه، %60 مجموعه داده ها به صورت مجموعه آموزشی جزء بندی شده اند، %20 به عنوان مجموعه اعتبار بخشی و %20 به عنوان مجموعه آزمون. شبکه های بسیاری آموزش داده شده اند و بهترین شبکه در مجموعه اعتباربخشی برای آزمون انتخاب شده اند؛ تابع آن در مجموعه آزمون برآورد شده است. میانگین نرخ دسته بندی، در چهار جنبه، به عنوان تخمین تابع تعمیم استفاده می شود. از آنجاییکه سرعت دسته بندی کلی مناسبترین ابزار برای اطلاعات نامتوازن نیست، ابزارهای دیگر اندازه گیری تابع تعمیم همچنین اجرا می شوند که شامل ابزار هندسی و ضریب کاپا می شود.



جدول 2. خلاصه ای از مجموعه های داده مورد استفاده در آزمایشها

نتایج مقایسه ای الگوریتمهای آموزشی مختلف در تمام مجموعه اطلاعات در جدول 1 نشان داده شده اند. الگوریتمهای آموزشی بهبود یافته و الگوریتمهای آموزشی استاندارد تقریبا نرخ دسته بندی یکسانی دارند. برای مثال، در مجموعه داده های هپاتیت، CRs ­متعلق به RPROP و Mod-RPROP به ترتیب %95.48 و %96.77 هستند. با این وجود، الگوریتمهای بهبود یافته ارزشهای بیشتر ابزار-g و ضریب کاپا نسبت به قرین های خود دارند. در مجموعه داده های هپاتیت، ارزشهای ابزار-g در RPROP و Mod-RPROP %94.90 و %97.03 هستند و ضریب کاپا RPROP و Mod-RPROP به ترتیب 0.878 و 0.919 است. این کشف اشاره می کند که الگوریتمهای آموزشی بهبود یافته نرخ دسته بندی خوبی را در همه دسته ها به نمایش می گذارد. شکل 1 میانگین CR s  را بوسیله دسته در چهار الگوریتم آموزشی نشان می دهد. نرخهای دسته بندی دسته مثبت بالا رفته است، برای مثال، %5.25 پیشرفت در مجموعه داده های کبد، %5 در مجموعه داده های هپاتیت، و %0.38 پیشرفت در مجموعه داده های دیابت.



شکل 1. میانگین نرخهای دسته بندی هر دسته بر روی همه الگوریتمهای آموزشی

# 4. نتیجه گیری

در این مقاله، یک روش جدید آموزشی برای شبکه های عصبی پیش خور در مجموعه دادهای نامتوازن که خوشه بندی نظارت نشده را با یادگیری نظارت شده ترکیب می کند ارائه شده است. روش پیشنهاد شده می تواند در الگوریتمهای آموزشی موجود به کار گرفته شود. نتایج آزمایش نشان می دهند که روش پیشنهاد شده می تواند به نحو موثری صحت دسته بندی دسته های اقلیت را ارتقا بخشد درحالیکه صحت عملکرد کلی دسته بندی را حفظ می کند.

# مراجع

[1] R. Alejo, V. Garcia, J. M. Sotoca, R. A. Mollineda, and

J. S. Senchez. *Improving the performance of the RBF*

*neural networks trained with imbalanced samples*, volume

4507 of *Lecture Notes in Computer Science*, pages

162–169. Springer-Verlag Berlin Heidelberg, 2007.

[2] V. L. Berardi and G. P. Zhang. The effect of misclassification

costs on neural network classifiers. *Decision*

*Sciences*, 30(3):659–683, 1999.

[3] R. Congalton, R. Oderwald, and R. Mead. Assessing

landsat classification accuracy using discrete multivariate

statistical techniques. *Photogrammetric Engineering*

*and Remote Sensing*, 49:1671–1678, 1983.

[4] X. Fu, L. Wang, K. S. Chua, and F. Chu. Training rbf

neural networks on unbalanced data. In *Neural Information*

*Processing*, volume 2, pages 1016– 1020, Inst.

of High Performance Comput., Singapore, 2002.

[5] M. Hagan and M. Menhaj. Training feedforward networks

with the marquardt algorithm. *IEEE Transactions*

*on Neural Networks*, 5:989–993, 1994.

[6] M. T. Hagan, H. B. Demuth, and M. H. Beale. *Neural*

*network design*. PWS Publishing, Boston, MA, 1996.

[7] Y. Lu, H. Guo, and L. Feldkamp. Robust neural learning

from unbalanced data samples. In *IEEE International*

*Joint Conference on Neural Networks, IEEE*

*World Congress on Computational Intelligence*, volume

3, pages 1816 – 1821, May 1998.

[8] Y. L. Murphey, H. Guo, and L. Feldkamp. Neural learning

from imbalanced data. In 117-128, editor, *Applied*

*Intelligence, special issue on Neural Networks and Applications*,

volume 21, 2004.

[9] Y. L. Murphey, H. Wang, G. Ou, and L. Feldkamp.

Oaho: an effective algorithm for multi-class learning

from imbalanced data. In *IEEE International Joint Conference*

*on Neural Networks*, pages 406–411, 2007.

[10] D. Newman, S. Hettich, C. L. Blake, and C. J. Merz.

UCI repository of machine learning databases, 1998.

[11] G. H. Nguyen, A. Bouzerdoum, and S. Phung. Efficient

supervised learning with reduced training examplars.

In *International Joint Conference on Neural Networks*,

2008.

[12] M. Riedmiller and H. Braun. A direct adaptive method

for faster backpropagation learning: The rprop algorithm.

In *IEEE International Conference on Neural Networks*,

volume 1, pages 586 – 591, 1993.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams.

Learning internal representations by error propagation.

In *Parallel distributed processing: explorations in the*

*microstructure of cognition*, volume I, pages 318 – 362.

Bradford Books, Cambridge, MA, 1986.

[14] K. Yoon and S. Kwek. A data reduction approach for

resolving the imbalanced data issue in functional genomics.

*Neural Comput and Applic*, 16:295–306, 2007.

[15] Z. Zhou and X. Liu. Training cost-sensitive neural

networks with methods addressing the class imbalance

problem. *IEEE Transactions on Knowledge and Data*