Kalina Kobus

Efektywne algorytmy dla wieloetykietowej klasyfikacji ekstremalnej

Streszczenie rozprawy doktorskiej

Promotor: dr hab. inż. Krzysztof Dembczyński

Poznań · 2020

Streszczenie

Uczenie maszynowe to dziedzina nauki z pogranicza informatyki, teorii informacji i statystyki. Rozwój uczenia maszynowego dotyczy zarówno jego aspektów teoretycznych, jak i praktycznych, a kolejne obszary zastosowania uczenia maszynowego otwierają nowe obszary badań teoretycznych. Wyróżnia się trzy główne obszary uczenia maszynowego: uczenie nadzorowane, uczenie nienadzorowane oraz uczenie ze wzmocnieniem. Niniejsza rozprawa dotyczy problemu klasyfikacji, będącego obiektem badań uczenia nadzorowanego.

W problemach klasyfikacji zadaniem jest poprawne wskazanie etykiety (lub zbioru etykiet) dla instancji, będącej reprezentacją pewnego obiektu, na podstawie jej cech. Poprawność tego wskazania określa się za pomocą funkcji straty. Aby rozwiązać problem klasyfikacji zazwyczaj trenuje się klasyfikator będący funkcją przypisującą instancjom etykiety na podstawie ich cech. Taki klasyfikator jest wynikiem wykonania algorytmu uczącego, działającego na danych treningowych, składających się z zaobserwowanych instancji z przypisanymi etykietami. Pożądaną własnością algorytmu uczącego jest to, że mając do dyspozycji coraz większą próbkę danych treningowych dostarcza on klasyfikator ze stratą coraz bliższą najniższej możliwej stracie, osiąganej przez tak zwany klasyfikator bayesowski. Taką własność określamy statystyczną zgodnością.

Klasyfikacja ekstremalna

W standardowych problemach klasyfikacji liczba istniejących etykiet jest nieduża. Jednak w wielu nowoczesnych obszarach zastosowania uczenia maszynowego mogą istnieć nawet miliony etykiet. Problemy z tak dużą liczą etykiet rozważane są w dziedzinie klasyfikacji ekstremalnej. Przykładowymi problemami klasyfikacji ekstremalnej są tagowanie dokumentów tekstowych [Dekel and Shamir, 2010], rekomendacja słów kluczowych dla reklam internetowych [Prabhu and Varma, 2014], rekomendacja wideo [Weston et al., 2013], czy predykcja kolejnego słowa w zdaniu [Mikolov et al., 2013]. Aby lep-

iej zrozumieć jak te problemy można przedstawić jako problem klasyfikacji przeanalizujmy następujące przykłady. W przypadku tagowania dokumentów tekst stanowi cechy, a kategorie to etykiety. W przypadku rekomendacji słów kluczowych dla reklam internetowych, cechy są tworzone na podstawie strony docelowej reklamy, a zapytania do wyszukiwarki stanowią etykiety. W przypadku rekomendacji wideo, użytkownik i wideo mogą być zamiennie używane jako cechy i etykiety. We wszystkich tych problemach liczba możliwych etykiet jest bardzo duża.

Celem klasyfikacji ekstremalnej jest predykcja określonej liczby k adekwatnych etykiet lub stworzenie ich rankingu dla danej instancji. Jakość predykcji często jest mierzona za pomocą precyzji na k-tym miejscu (*precision@k*), zdefiniowanej jako udział pozytywnych etykiet wśród k przewidzianych, czy NDCG@k, będącej miarą typową dla problemów rangowania, która przypisuje poprawnie przewidzianym etykietom zysk malejący wraz z rangą etykiety. Innymi miarami używanymi w klasyfikacji ekstremalnej są czułość na k-tym miejscu i makro-uśredniana miara F_1.

Z uwagi na liczbę istniejących etykiet, efektywność obliczeniowa algorytmów uczenia i predykcji w klasyfikacji ekstremalnej odgrywa większą rolę niż w problemach mniejszej skali. W klasyfikacji ekstremalnej nawet algorytmy skalujące się liniowo z liczbą etykiet mogą być zbyt wolne aby mogły być użyteczne. Przykładem takiego algorytmu jest standardowe rozwiązanie problemów wieloetykietowych, polegające na nauczeniu niezależnego klasyfikatora dla każdej etykiety z osobna. Takie podejście nazywane jest 1-VS-ALL (jeden przeciwko wszystkim/innym). Charakteryzuje się ono liniową wobec liczby etykiet złożonością czasową i pamięciową. Taka złożoność jest zbyt duża w wielu praktycznych zastosowaniach. Przykładowo, rozważmy problem z 10^6 etykietami. Załóżmy, że pojedynczy klasyfikator można wytrenować w jedną sekundę. W takim przypadku trening 10^6 klasyfikatorów zająłby ponad 11 dni. Zauważmy również, że w sytuacji gdy w zbiorze treningowym znajduje się bardzo wiele cech i obserwacji, zakładany czas treningu wynoszący jedną sekundę może być zdecydowanie zbyt niski. Również predykcja z użyciem tego podejścia jest czasochłonna, ponieważ wymaga ewaluacji 10^6 klasyfikatorów. Ponadto, gdy rozważymy również rozmiar modelu, przy założeniu istnienia 10^5 cech i użycia gęstych klasyfikatorów liniowych, łatwo uzyskujemy rozmiary modelu rzędu setek gigabajtów. Widzimy, że konieczne jest stworzenie bardziej zaawansowanych rozwiązań, cechujących się dobrą jakością predykcji i niższą od liniowej złożonością.

Zauważmy, że klasyfikacja ekstremalna stawia również inne, nie tylko obliczeniowe, wyzwania nieobecne w standardowych problemach uczenia. Przykładowo, dla wielu etykiet w zbiorach uczących znajduje się bardzo mało obserwacji. Spotyka się również problemy, w których etykiety nie posiadają żadnych obserwacji w zbiorze uczącym. Jest to tak zwany problem uczenia z zera próbek. Ponadto, dane treningowe są zazwyczaj niskiej jakości, ponieważ jest niemożliwe aby ręcznie zweryfikować wszystkie możliwe etykiety nawet dla pojedynczej obserwacji. Często dane uczące są uzyskiwane nie wprost, co

prowadzi do całego spektrum problemów. W niniejszej pracy skupiamy się jednak przede wszystkim na jakości predykcji oraz efektywności obliczeń, a nie na wyżej wymienionych wyzwaniach.

Istniejące metody

W dziedzinie klasyfikacji ekstremalnej zaproponowanych zostało wiele metod mających na celu osiągnięcie wysokiej jakości predykcji przy niskich czasach uczenia i predykcji. Standardowym podejściem cechującym się dużą wydajnością i jakością predykcji są drzewa decyzyjne. Jednakże w klasyfikacji ekstremalnej nie można zastosować standardowych drzew decyzyjnych właśnie ze względu na ich koszty obliczeniowe, wysokie w przypadku problemów ekstremalnych [Agrawal et al., 2013], wynikające z konieczności obliczenia kryterium podziału wierzchołka. Algorytm FASTXML [Prabhu and Varma, 2014] redukuje ten koszt przez użycie klasyfikatorów liniowych we wierzchołkach decyzyjnych drzewa. Te klasyfikatory są wynikiem naprzemiennej optymalizacji specyficznej wielokryterialnej funkcji celu. FASTXML używa wielu drzew w celu poprawy jakości predykcji. Idea FASTXML została rozszerzona w PFASTREXML [Jain et al., 2016] i CRAFTML [Siblini et al., 2018], które modyfikują optymalizowane kryterium lub sposób jego optymalizacji. W pełni przyrostową metodę budowy drzewa dla problemów wieloklasowych rozważono w [Choromanska and Langford, 2015], proponując LOMTREE. Z kolei w [Majzoubi and Choromanska, 2019] zaproponowano algorytm LDSM dla problemów wieloklasowych, budujący drzewo wierzchołek po wierzchołku w częściowo przyrostowy sposób. Przez długi czas metody oparte na drzewach decyzyjnych osiągały najlepsze wyniki pod względem jakości predykcji i wydajności obliczeniowej.

Innym podejściem stosowanym w klasyfikacji ekstremalnej są zanurzenia. Metody tego typu redukują oryginalną przestrzeń wyjść do przestrzeni o mniejszej liczbie wymiarów i tworzą modele regresyjne w tej zredukowanej przestrzeni [Tai and Lin, 2012], a następnie transformują predykcje z przestrzeni zredukowanej do oryginalnej. Różnią się one sposobem wykonania tej kompresji i dekompresji. W [Bhatia et al., 2015] zaproponowano rzadkie lokalne zanurzenia (*sparse local embeddings*, SLEEC) używające klasyfikatora k najbliższych sąsiadów w przestrzeni zredukowanej do dekompresji predykcji. Jakość predykcji algorytmu SLEEC na zbiorach porównawczych jest porównywalna z jakością innych algorytmów. Jednak istotną wadą tej metody są duże rozmiary modeli i długie czasy treningu i predykcji. ANNEXML [Tagami, 2017] poprawia efektywność predykcji przez użycie grafu k najbliższych sąsiadów. Ostatnio, GLAS [Guo et al., 2019] osiągnął konkurencyjną jakość predykcji i zredukował czas potrzebny na ich uzyskanie poprzez użycie informacji o współwystąpieniach etykiet oraz szybkich metod wyszukiwania największego iloczynu skalarnego.

Jakość predykcji metody 1-vs-All długo była uważana za trudną do osiągnięcia metodami mniej kosztownymi obliczeniowo. Tak zwane sprytne metody 1-vs-All trenują jeden binarny klasyfikator dla każdej etykiety, ale redukują koszty obliczeniowe przez użycie rozproszonych obliczeń i ucinania wag (DiSMEC, [Babbar and Schölkopf, 2017]), odpowiednich metod optymalizacji i predykcji (PD-Sparse, [Yen et al., 2016]; PPD-Sparse, [Yen et al., 2017]), czy obu (ProXML, [Babbar and Schölkopf, 2019]). Te metody osiągają bardzo wysoka jakość predykcji, jednak ich czasy obliczeń i treningu nadal pozostają znacznie dłuższe niż czasy osiągane przez inne metody.

Model probabilistycznych drzew etykiet (*probabilistic label trees*, PLT), będący pierwszą metodą wieloetykietowej klasyfikacji ekstremalnej opartą na drzewach etykiet, został zaproponowany w [Jasinska et al., 2016]. Drzewa etykiet umożliwiają efektywne przybliżenie modelu 1-vs-All, przy krótszych czasach predykcji i treningu. Drzewa etykiet różnią się znacznie od drzew decyzyjnych, ponieważ w drzewie etykiet każda ścieżka odpowiada dokładnie jednej etykiecie, a nie fragmentowi przestrzeni cech. Podejście oparte na modelu PLT zostało później wykorzystane w takich algorytmach jak extremeText Wydmuch et al. [2018], Parabel [Prabhu et al., 2018], Bonsai Tree [Babbar and Schölkopf, 2019], czy AttentionXML [You et al., 2019]. Powyższe algorytmy należą do najbardziej popularnych i uznanych algorytmów wieloetykietowej klasyfikacji ekstremalnej.

Metody oparte na uczeniu głębokim również zostały zastosowane do klasyfikacji ekstremalnej. Metody te, zastosowane do danych tekstowych, używają oryginalnej reprezentacji danych, podczas gdy pozostałe metody używają rzadkich reprezentacji tekstu. Z tego względu trudno wprost porównać jakość predykcji tych metod do jakości predykcji pozostałych. Również ze względu na użycie obliczeń na kartach graficznych, czasy treningu i predykcji nie są wprost porównywalne. Pierwszą metodą tego typu w wieloetykietowej klasyfikacji ekstremalnej jest XML-CNN [Liu et al., 2017]. Nie dość, że osiąga on gorszą jakość predykcji niż inne metody, to charakteryzuje się bardzo długimi czasami treningu i predykcji. Wspomniany wcześniej AttentionXML [You et al., 2019] używa płytkiego PLT i specyficznego mechanizmu wieloetykietowej uwagi. Również X-BERT [Chang et al., 2019] może być traktowany jako PLT z klasyfikatorami we wierzchołkach wewnętrznych opartymi o sieci głębokie, a dokładniej o przed-trenowany model BERT [Devlin et al., 2018], oraz liniowe w liściach.

Wiele metod zostało zaproponowanych w celu rozwiązania problemów stawianych przez wieloetykietową klasyfikację ekstremalną. Jakkolwiek te metody pozwalają osiągnąć wysoką jakość predykcji szybciej niż naiwnie zaaplikowana metoda 1-vs-All, to niewiele z nich zostało przeanalizowanych pod względem statystycznej zgodności ze względu na optymalizowane miary oceny, czy ze względu na złożoność obliczeniową.

Motywacje

Motywację zaproponowanych probabilistycznych drzew etykiet stanowi prosta obserwacja: okazuje się, że optymalne predykcje, czyli tak zwane klasyfikatory bayesowskie, ze względu na precyzję na k-tym miejscu i inne popularne metryki, można określić za pomocą prawdopodobieństw warunkowych (ze względu na cechy) etykiet. A zatem, estymacja tych prawdopodobieństw i użycie odpowiedniej reguły decyzyjnej mogłyby zagwarantować statystyczną zgodność zaproponowanej metody. Z tej perspektywy, problem wieloetykietowej klasyfikacji ekstremalnej wydaje się problemem efektywnej estymacji prawdopodobieństw i efektywnego wnioskowania.

Taka efektywność estymacji prawdopodobieństw warunkowych etykiet może być uzyskana przez zorganizowanie etykiet w drzewo, w którym każdej etykiecie odpowiada jeden liść, czyli w tak zwane drzewo etykiet. Tego typu metody są używane w klasyfikacji wieloklasowej. Przykładem jest hierarchiczny softmaks [Morin and Bengio, 2005], używany w sieciach głębokich, między innymi w przetwarzaniu języka naturalnego [Mikolov et al., 2013]. Co ciekawe, podobne algorytmy były zaproponowane niezależnie w innych dziedzinach. W statystyce znane są jako *nested dichotomies* [Fox, 1997], w wieloklasowej regresji jako drzewa prawdopodobieństw warunkowych [Beygelzimer et al., 2009], a w rozpoznawaniu wzorców jako wieloetapowe klasyfikatory [Kurzynski, 1988]. Jednakże to podejście nie zostało wcześniej zastosowane w wieloetykietowej klasyfikacji ekstremalnej.

Cel i kontrybucje

Ze względu na przedstawione wcześniej motywacje, sformułowano następującą hipotezę rozprawy:

> Istnieje klasa statystycznie zgodnych algorytmów uczenia dla wieloetykietowej klasyfikacji ekstremalnej charakteryzujących się podliniową złożonością obliczeniową względem liczby etykiet.

Poniżej opisany jest główny wkład przedstawiony w rozprawie.

Postaci klasyfikatorów bayesowskich

W rozprawie dokonujemy przeglądu miar oceny jakości klasyfikacji używanych w ekstremalnej klasyfikacji wieloetykietowej. Dowodzimy, że klasyfikatorem bayesowskim dla precyzji na k-tym miejscu jest wskazanie k etykiet z najwyższym prawdopodobieństwem warunkowym. Podobnie pokazujemy postaci klasyfikatora bayesowskiego dla miar DCG@k oraz

NDCG@k. Dodatkowo na podstawie literatury omawiamy postaci klasyfikatora bayesowskiego dla uogólnionych miar oceny jakości klasyfikacji [Kotłowski and Dembczyński, 2017] oraz czułości na @k-tym miejscu [Menon et al., 2019]. W ten sposób pokazujemy, dla których miar oceny jakości klasyfikacji optymalne predykcje mogą być określone na podstawie warunkowych prawdopodobieństw etykiet.

Model PLT

Proponujemy i opisujemy model probabilistycznych drzew etykiet (en. *probabilistic label trees*, PLT). Probabilistyczne drzewa etykiet używają drzewa etykiet do rozkładu prawdopodobieństwa warunkowego etykiet poprzez zastosowanie reguły łańcuchowej wzdłuż ścieżki od korzenia drzewa do liścia odpowiadającego etykiecie. W ten sposób redukują one oryginalny problem wieloetykietowy do wielu problemów klasyfikacji (estymacji) binarnej. Z tego punktu widzenia PLT jest przedstawicielem *redukcji uczenia* [Beygelzimer et al., 2016]. PLT używa probabilistycznych klasyfikatorów binarnych we wszystkich wierzchołkach drzewa do estymacji odpowiednich czynników, będących prawdopodobieństwami warunkowymi. Iloczyn estymat prawdopodobieństwa na ścieżce od korzenia do liścia jest estymatą prawdopodobieństwa warunkowego etykiety odpowiadającej liściowi. Do efektywnej predykcji PLT używa odpowiednich procedur opartych na przeszukiwaniu drzewa.

Rozważamy dwa sposoby uczenia PLT: trening wsadowy lub przyrostowy przy danej strukturze drzewa etykiet, oraz trening w pełni przyrostowy, w którym drzewo jest konstruowane jednocześnie z treningiem klasyfikatorów. Dowodzimy specyficzną tożsamość klasyfikatora PLT nauczonego przyrostowo oraz w pełni przyrostowo. Analizujemy trzy sposoby przeszukiwania drzewa w celu predykcji. Pierwszy odnajduje wszystkie etykiety o estymowanym prawdopodobieństwie warunkowym przekraczającym wskazany próg. Drugi z nich, oparty na przeszukiwaniu ze strategią jednolitego kosztu odnajduje wskazaną liczbę etykiet o najwyższych estymatach prawdopodobieństwa warunkowego. Trzeci, oparty na przeszukiwaniu wiązkowym, odnajduje przybliżone etykiety z najwyższą estymatą.

Zgodność i ograniczenia na żal

Wyniki teoretyczne związane z PLT dotyczą jego zgodności ze względu na wspomniane wcześniej miary oceny jakości klasyfikacji. Zgodność wykazujemy zgodnie z metodyką redukcji uczenia [Beygelzimer et al., 2016], ograniczając błąd L_1 estymacji prawdopodobieństw warunkowych etykiet za pomocą funkcji żalu klasyfikatorów wierzchołkowych, wyrażonego ze względu na silnie właściwy złożony błąd zastępczy. W tym celu wpierw ograniczamy błąd L_1 estymacji prawdopodobieństwa warunkowego etykiety za pomocą błędów L_1 estymacji prawdopodobieństw warunkowych związanych z wierzchołkami na ścieżce od korzenia do liścia odpowiadającego etykiecie. Następnie wyrażamy

błąd L_1 każdego wierzchołka za pomocą silnie właściwego złożonego błędu zastępczego [Agarwal, 2014]. To pozwala nam połączyć żal klasyfikatorów wierzchołkowych z błędem L_1 estymacji prawdopodobieństw warunkowych etykiet. Ten wynik stanowi podstawę kolejnych ograniczeń ze względu na różne miary oceny jakości klasyfikacji, dla których optymalne predykcje można określić za pomocą prawdopodobieństw warunkowych etykiet.

Za pomocą wyprowadzonego ograniczenia na błąd L_1, pokazujemy ograniczenia żalu ze względu na uogólnione miary oceny jakości klasyfikacji. Do tej klasy funkcji należy między innymi strata Hamminga oraz mikro- i makro- uśredniana miara F_1. Wyprowadzone ograniczenia bazują na wynikach z [Kotłowski and Dembczyński, 2017] dotyczących metody 1-VS-ALL. Następnie rozważamy precyzję na k-tym miejscu. Definiujemy żal oraz ograniczamy go za pomocą błędu L_1 estymacji prawdopodobieństw etykiet. W ten sposób pokazujemy, że PLT używające dokładnej metody predykcji k etykiet z najwyższym estymowanym prawdopodobieństwem warunkowym, jest dostosowane do optymalizacji precyzji na k-tym miejscu. W podobny sposób analizujemy miarę DCG@k, pokazując analogiczne wyniki dla tej miary.

Analizujemy także związek pomiędzy PLT i hierarchicznym softmaksem. Pokazujemy, że PLT jest poprawnym uogólnieniem hierarchicznego softmaksu do problemów wieloetykietowych. Oznacza to, że dla danych wieloklasowych model PLT redukuje się do modelu hierarchicznego softmaksu. Ponadto pokazujemy, że inna popularna metoda uogólnienia hierarchicznego softmaksu, stosująca heurystykę wyboru jednej etykiety, stosowana przykładowo w FASTTEXT [Joulin et al., 2017] i LEARNED TREE [Jernite et al., 2017], nie jest zgodna ze względu na estymację prawdopodobieństw warunkowych etykiet przy błędzie L_1 i ze względu na precyzję na k-tym miejscu.

Złożoność obliczeniowa algorytmów uczących i predykcyjnych

Analizujemy złożoność obliczeniową algorytmów PLT służących do uczenia i predykcji. Pokazujemy, że uczenie klasyfikatorów PLT przy określonej strukturze drzewa, przy pewnych dodatkowych założeniach dotyczących struktury drzewa oraz maksymalnej liczby pozytywnych etykiet na obserwację, może być wykonany w czasie logarytmicznym ze względu na liczbę etykiet. Ponadto, pokazujemy, że przy dodatkowych założeniach dotyczących estymat prawdopodobieństw, również czas predykcji jest logarytmiczny, lub podliniowy, we względu na liczbę etykiet. Wyniki te nie są trywialne, ponieważ metody przeszukiwania drzewa na których bazuje predykcja w najgorszym przypadku mogą przeszukać całe drzewo, co prowadziłoby do złożoności liniowej.

Ocena empiryczna

Poza wynikami teoretycznymi analizujemy istniejące implementacje ogólnego schematu PLT, takie jak XMLC-PLT [Jasinska et al., 2016], PLT-VW[1], PARABEL [Prabhu et al., 2018], BONSAI TREE [Khandagale et al., 2019], EXTREMETEXT [Wydmuch et al., 2018], ATTENTIONXML [You et al., 2019], oraz NAPKINXC [Jasinska-Kobus et al., 2020a]. W analizie koncentrujemy się na możliwych sposobach implementacji ze względu na reprezentację cech i modeli oraz metody treningu i predykcji.

W części eksperymentalnej przede wszystkim koncentrujemy się na implementacjach NAPKINXC oraz PARABEL. Za ich pomocą analizujemy różne instancje modelu PLT i porównujemy uzyskiwane przez nie wyniki do powszechnie uznanych metod bazujących na drzewach decyzyjnych oraz metodach 1-VS-ALL. Empirycznie wykazujemy, że PLT jest konkurencyjne wobec najlepszych metod, i uzyskuje najwyższe wartości precyzji na pierwszym miejscu na większości zbiorów porównawczych, będąc jednocześnie trzy rzędy wielkości szybsze od metod 1-VS-ALL.

Przegląd prac stanowiących podstawę rozprawy

Poniżej dokonujemy przeglądu prac dotyczących PLT stanowiących podstawę rozprawy. Pierwsza praca dotycząca PLT [Jasinska and Dembczyński, 2015] została przedstawiona na warsztacie *Extreme Classification Workshop* przy konferencji ICML 2015. Ten artykuł wprowadzał model PLT oraz inny model, BRT, również będący drzewem etykiet. PLT wraz z prostymi metodami treningu i predykcji zostało następnie opublikowane w [Jasinska et al., 2016]. Ta praca dotyczy użycia PLT dla optymalizacji precyzji na k-tym miejscu oraz makrouśrednianej miary F_1. Następnie w [Jasinska, 2018] zaproponowano wsadowy wariant predykcji używającej przeszukiwania ze strategią jednolitego kosztu. Następnie w [Wydmuch et al., 2018] przeanalizowano błąd L_1 estymacji oraz ograniczono żal ze względu na precyzję na k-tym miejscu. Złożoność obliczeniowa PLT została przeanalizowana w [Busa-Fekete et al., 2019]. Niniejsza rozprawa zawiera część wyników z tej pracy, dotyczących ograniczeń kosztów treningu i predykcji. W pełni przyrostowe PLT zostało zaproponowane w roku 2016, a opublikowane w [Jasinska-Kobus et al., 2020b,c]. Większość wyników teoretycznych przedstawionych w niniejszej rozprawie znajduje się w [Jasinska-Kobus et al., 2020a]. Niniejsza rozprawa zawiera również rezultaty niezależne od PLT. W [Jasinska and Karampatziakis, 2016] przedstawiono inny algorytm klasyfikacji ekstremalnej nazwany LTLS. Wyniki dotyczące NDCG@k zostały przedstawione w [Jasinska and Dembczyński, 2018].

[1] https://github.com/VowpalWabbit/vowpal_wabbit

Bibliografia

S. Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15:1653–1674, 2014.

R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the 22nd International Conference on World Wide Web*, page 13–24, New York, NY, USA, 2013. Association for Computing Machinery.

R. Babbar and B. Schölkopf. Dismec: Distributed sparse machines for extreme multi-label classification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, page 721–729, New York, NY, USA, 2017. Association for Computing Machinery.

R. Babbar and B. Schölkopf. Data scarcity, robustness and extreme multi-label classification. *Machine Learning, Special Issue of the ECML PKDD 2019 journal Track*, 108, 2019.

A. Beygelzimer, J. Langford, and P. Ravikumar. Error-correcting tournaments. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory*, page 247–262, Berlin, Heidelberg, 2009. Springer-Verlag.

A. Beygelzimer, H. Daumé, J. Langford, and P. Mineiro. Learning reductions that really work. *Proceedings of the IEEE*, 104:136–147, 2016.

K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain. Sparse local embeddings for extreme multi-label classification. In *Advances in Neural Information Processing Systems 28*, pages 730–738. Curran Associates, Inc., 2015.

R. Busa-Fekete, K. Dembczynski, A. Golovnev, K. Jasinska, M. Kuznetsov, M. Sviridenko, and C. Xu. On the computational complexity of the probabilistic label tree algorithms. *CoRR*, abs/1906.00294, 2019.

W. Chang, H. Yu, K. Zhong, Y. Yang, and I. S. Dhillon. A modular deep learning approach for extreme multi-label text classification. *CoRR*, abs/1905.02331, 2019.

A. E. Choromanska and J. Langford. Logarithmic time online multiclass prediction. In *Advances in Neural Information Processing Systems 28*, pages 55–63. Curran Associates, Inc., 2015.

O. Dekel and O. Shamir. Multiclass-multilabel classification with more classes than examples. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 137–144, Chia Laguna Resort, Sardinia, Italy, 2010. PMLR.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

J. Fox. *Applied regression analysis, linear models, and related methods*. Sage, 1997.

C. Guo, A. Mousavi, X. Wu, D. N. Holtmann-Rice, S. Kale, S. Reddi, and S. Kumar. Breaking the glass ceiling for embedding-based classifiers for large output spaces. In *Advances in Neural Information Processing Systems 32*, pages 4943–4953. Curran Associates, Inc., 2019.

H. Jain, Y. Prabhu, and M. Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 935–944, New York, NY, USA, 2016. Association for Computing Machinery.

K. Jasinska. Efficient exact batch prediction for label trees. In *Extreme Multilabel Classification for Social Media at The Web Conference*, 2018.

K. Jasinska and K. Dembczyński. Consistent label tree classifiers for extreme multi-label classification. In *The ICML Workshop on Extreme Classification*, 2015.

K. Jasinska and K. Dembczyński. Bayes optimal prediction for ndcg@k in extreme amulti-label classification. In *From Multiple Criteria Decision Aid to Preference Learning Workshop*, 2018.

K. Jasinska and N. Karampatziakis. Log-time and log-space extreme classification. *CoRR*, abs/1611.01964, 2016.

K. Jasinska, K. Dembczynski, R. Busa-Fekete, K. Pfannschmidt, T. Klerx, and E. Hullermeier. Extreme f-measure maximization using sparse probability estimates. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1435–1444, New York, USA, 2016. PMLR.

K. Jasinska-Kobus, M. Wydmuch, K. Dembczyński, M. Kuznetsov, and R. Busa-Fekete. Probabilistic label trees for extreme multi-label classification. *Journal of Machine Learning Research (in review)*, 2020a.

K. Jasinska-Kobus, M. Wydmuch, D. Thiruvenkatachari, and K. Dembczyński. Online probabilistic label trees. *CoRR*, abs/2007.04451, 2020b.

K. Jasinska-Kobus, M. Wydmuch, D. Thiruvenkatachari, and K. Dembczyński. Online probabilistic label trees. *AISTATS 2020 (in review)*, 2020c.

Y. Jernite, A. Choromanska, and D. Sontag. Simultaneous learning of trees and representations for extreme classification and density estimation. In *Proceedings of the 34th International Conference on Machine Learning - volume 70*, page 1665–1674. JMLR.org, 2017.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: volume 2, Short Papers*, pages 427–431, Valencia, Spain, 2017. Association for Computational Linguistics.

S. Khandagale, H. Xiao, and R. Babbar. Bonsai - diverse and shallow trees for extreme multi-label classification. *CoRR*, abs/1904.08249, 2019.

W. Kotłowski and K. Dembczyński. Surrogate regret bounds for generalized classification performance metrics. *Machine Learning*, 10:549–572, 2017.

M. Kurzynski. On the multistage bayes classifier. *Pattern Recognition*, 21:355–365, 1988.

J. Liu, W.-C. Chang, Y. Wu, and Y. Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 115–124, New York, NY, USA, 2017. Association for Computing Machinery.

M. Majzoubi and A. Choromanska. Ldsm: Logarithm-depth streaming multi-label decision trees. *CoRR*, abs/1905.10428, 2019.

A. K. Menon, A. S. Rawat, S. Reddi, and S. Kumar. Multilabel reductions: what is my loss optimising? In *Advances in Neural Information Processing Systems 32*, pages 10600–10611. Curran Associates, Inc., 2019.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 246–252. Society for Artificial Intelligence and Statistics, 2005.

Y. Prabhu and M. Varma. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 263–272, New York, NY, USA, 2014. Association for Computing Machinery.

Y. Prabhu, A. Kag, S. Harsola, R. Agrawal, and M. Varma. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference*, page 993–1002, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

W. Siblini, P. Kuntz, and F. Meyer. CRAFTML, an efficient clustering-based random forest for extreme multi-label learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4664–4673, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

Y. Tagami. Annexml: Approximate nearest neighbor search for extreme multi-label classification. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 455–464, New York, NY, USA, 2017. Association for Computing Machinery.

F. Tai and H.-T. Lin. Multilabel classification with principal label space transformation. *Neural Computation*, 24:2508–2542, 2012.

J. Weston, A. Makadia, and H. Yee. Label partitioning for sublinear ranking. In *Proceedings of the 30th International Conference on Machine Learning*, pages 181–189, Atlanta, Georgia, USA, 2013. PMLR.

M. Wydmuch, K. Jasinska, M. Kuznetsov, R. Busa-Fekete, and K. Dembczynski. A no-regret generalization of hierarchical softmax to extreme multi-label classification. In *Advances in Neural Information Processing Systems 31*, pages 6355–6366. Curran Associates, Inc., 2018.

I. E. Yen, X. Huang, W. Dai, P. Ravikumar, I. Dhillon, and E. Xing. Ppdsparse: A parallel primal-dual sparse method for extreme classification. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 545–553. Association for Computing Machinery, 2017.

I. E.-H. Yen, X. Huang, P. Ravikumar, K. Zhong, and I. Dhillon. Pd-sparse : A primal and dual sparse approach to extreme multiclass and multilabel classification. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 3069–3077, New York, New York, USA, 2016. PMLR.

R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *Advances in Neural Information Processing Systems 32*, pages 5820–5830. Curran Associates, Inc., 2019.