

ՀՀ ԳԱԱ ԻՆՖՈՐՄԱՏԻԿԱՅԻ ԵՎ ԱՎՏՈՄԱՏԱՑՄԱՆ ՊՐՈԲԼԵՄՆԵՐԻ ԻՆՏԻՏՈՒՏ

Օղլուկյան Արթուր Կարապետի

ՏԵՂԱԲԱԾԽՎԱԾ ԵՎ ՁՈՒԳԱՅԵՐԻ ՄԵԶԵՆԱՅԱԿԱՆ ՈՒՄՈՒՑՄԱՆ ԼՈՒԾՈՒՄՆԵՐԻ
ՆԱԽԱԳԾՈՒՄ

Ե.13.04 - «Հաշվողական մեքենաների, համալիրների, համակարգերի և ցանցերի
մաթեմատիկական և ծրագրային ապահովում» մասնագիտությամբ տեխնիկական
գիտությունների թեկնածուի գիտական աստիճանի հայցման ատենախոսության
սեղմագիր

Երևան 2025

INSTITUTE FOR INFORMATICS AND AUTOMATION PROBLEMS OF THE NAS RA

Oghlukyan Artur Karapet

DESIGNING DISTRIBUTED AND PARALLEL MACHINE LEARNING SOLUTIONS

An abstract of dissertation for obtaining a Ph.D. degree in Technical Sciences on specialty
05.13.04 - "Mathematical and Software Support of Computers, Complexes and Computer
Networks"

Yerevan 2025

Ատենախոսության թեման հաստատվել է ՀՀ ԳԱԱ Ինֆորմատիկայի և
Ավտոմատացման Պրոբլեմների Ինստիտուտում

Գիտական ղեկավար՝

Ֆ.մ.գ.դ., ՀՀ ԳԱԱ թղթ. Անդամ
Լ. Յ. Ասլանյան

Պաշտոնական ընդդիմախոսներ՝

տ.գ.դ. Յ. Վ. Ասցատրյան
տ.գ.թ. Ս. Վ. Գրիգորյան

Առաջատար կազմակերպություն՝ Հայաստանի Ազգային Պոլիտեխնիկական
Համալսարան (ՀԱՊՀ)

Ատենախոսության պաշտպանությունը կկայանա 2025թ. հուլիսի 14-ին, ժ. 14:00-ին
ԳԱԱ ԻԱՊԻ գործող 037 մասնագիտական խորհրդի նիստում հետևյալ հասցեով՝
Երևան, 0014, Պ. Սևակի 1:

Ատենախոսությանը կարելի է ծանոթանալ **ՀՀ ԳԱԱ ԻԱՊԻ** գրադարանում:

Սեղմագիրն առաքված է 2025թ. հունիսի 14-ին:

Մասնագիտական խորհրդի գիտական
քարտուղար՝ Տ.Մ.գ.դ.

Մ. Է. Հարությունյան

The topic of the dissertation was approved at the Institute for Informatics and Automation Problems of NAS RA

Scientific Supervisor: D.Ph.M.S., NAS RA correspondent member - L. H. Aslanyan

Official Opponents:

D.tech.s. H. V. Astsatryan
Ph.D.tech.s: S. V. Grigoryan

Leading Organization: National Polytechnic University of Armenia (NPUA)

The dissertation defence will take place on 14 of July, 2025, at 14:00 at the Specialized Council 037 «Informatics» at the Institute of Informatics and Automation Problems of NAS RA. Address: Yerevan, 0014, P. Sevak 1.

The Dissertation is available at the library of IIAP NAS RA.

The abstract was delivered on June 14th 2025.

Scientific Secretary of the Specialized Council,

D.Ph.M.S. M. E. Haroutunian

Ատենախոսության ընդհանուր բնութագիր

Թեմայի արդիականությունը

Արհեստական Բանականության և մեծ տվյալների դարաշրջանում մեքենայական ուսուցման ալգորիթմների արդյունավետ կիրառությունը տարբեր բաշխումների ենթակա տվյալների վրա հանդիսանում է կենսական խնդիր: Հատկապես սա կարևոր է PAC (Probably Approximately Correct) ուսուցման տեսական մոդելի շրջանակներում, որտեղ ալգորիթմները պետք է ապահովեն կարգավորելի ճշգրտություն և վստահություն կամայական բաշխումների պարագայում: PAC մոդելը ֆորմալացնում է մեքենայական ուսուցման հիմնական գաղափարները և սահմանում է այն պայմանները, որոնց դեպքում ալգորիթմը կարող է արդյունավետ սովորել:

Boosting ալգորիթմները, մասնավորապես AdaBoost-ը, հանդիսանում են PAC ուսուցման խնդիրների մասնակի լուծում՝ թույլ տալով համակցել թույլ դասակարգիչների հավաքածուները և ստեղծել ուժեղ դասակարգիչ տեսականորեն երաշխավորված կատարողականությամբ: Շափիրեի և Ֆրեյնդի տեսական արդյունքները ցույց են տվել, որ եթե գոյություն ունի թույլ սովորող ալգորիթմներ, որոնց արդյունավետությունը գերազանցում է պատահական գուշակումը, ապա boosting-ը կարող է կառուցել բարձր ճշգրտությամբ ուժեղ դասակարգիչ:

Սակայն առկա boosting մեթոդների հիմնական սահմանափակումը դրանց հաջորդական բնույթն է՝ յուրաքանչյուր թույլ մոդելը պետք է սովորեցվի նախորդների հիման վրա, ինչը դժվարացնում է մեծածավալ տվյալների արդյունավետ մշակումը: Այս սահմանափակումը հատկապես կրիտիկական է դառնում ժամանակակից կիրառություններում, որտեղ տվյալների ծավալները հասնում են պետաբայթների, ինչը պահանջում է ժամանակատար մշակում: Բացի այդ, ավանդական AdaBoost-ի հաղորդակցական բարդությունը և համաժամեցման պահանջները բախվածության պարագայում դարձնում են այն ոչ պրակտիկ մեծ համակարգերի համար, որտեղ ցանցային ուղացումները և հանգույցների անկանխատեսելի վարքագիծը կարող են խիստ ազդել ալգորիթմի զուգամիտման վրա:

Աշխատանքի նպատակն ու խնդիրները

Հիմնական նպատակ: AdaBoost դասի ալգորիթմներիի արդյունավետ բաշխված կառուցվածքով իրականացման մշակումն է՝ օգտագործելով բաշխվածության և զուգահեռացման ծրագրային տեխնոլոգիաներ, օրինակ MPI (Message Passing Interface) տեխնոլոգիան և ադապտիվ հաղորդակցական մեխանիզմներ:

Խնդիրները:

1. Բաշխված AdaBoost ալգորիթմների տեսական վերլուծություն
2. MPI-հիմնված ճարտարապետության ներդրում
3. Ադապտիվ հաղորդակցական ժամանակացույցի ներդրում
4. Բուֆերային մեխանիզմների իրականացում տեղական մոդելների կառավարման համար
5. Փորձարարական գնահատման իրականացում

Հետազոտման օբյեկտը

Հետազոտության օբյեկտն է AdaBoost boosting ալգորիթմի բաշխված իրականացումը՝ կենտրոնանալով հաղորդակցական արդյունավետության և մասշտաբավորելիության բարելավման վրա: Հետազոտությունը ուսումնասիրում է ավանդական AdaBoost ալգորիթմի առանձնահատկությունները և սահմանափակումները բաշխված միջավայրերում, որտեղ տվյալները բաժանված են մի քանի հաշվողական հանգույցների միջև:

Մասնավորապես, օբյեկտը ներառում է AdaBoost-ի հաջորդական բնույթի վերափոխումը բաշխված մոդելի, որտեղ յուրաքանչյուր հանգույց կարող է անկախ մշակել տվյալների իր հատվածը՝ պարբերաբար համաժամեցվելով մյուս հանգույցների հետ: Հետազոտության կենտրոնում է նաև թույլ դասակարգիչների տեղական ուսուցման և գլոբալ ինտեգրման մեխանիզմների օպտիմալացումը, ինչպես նաև հաղորդակցական ծախսերի նվազեցումը՝ պահպանելով տեսական երաշխիքները:

Հետազոտության շրջանակներում ուսումնասիրվում են ադապտիվ համաժամեցման մեթոդները, բուֆերային մեխանիզմները և MPI-հիմնված բաշխված ճարտարապետությունները՝ նպատակ ունենալով ստեղծել մասշտաբավորվող լուծում, որը կարող է արդյունավետ աշխատել տարբեր չափերի կլաստերներում և հարմարվել տարբեր ցանցային պայմանների:

Հետազոտման մեթոդները

- Մեքենայական ուսուցման տեսական մեթոդներ
- Բաշխված համակարգերի նախագծում
- MPI (Message Passing Interface) ծրագրավորում
- Գործնական փորձարկում supercomputing միջավայրերում
- Համեմատական վերլուծություն

1. **Ադապտիվ համաժամեցման մեխանիզմի մշակում.** Նորարարական մոտեցում համաժամեցման միջակայքի դինամիկ կարգավորման համար՝ հիմնվելով սխալի փոփոխման վրա
2. **ԲուՖերային ռազմավարության ներդրում.** լուրջ/տեղական մոդելների արդյունավետ կուտակում և կառավարում
3. **Սխալի չափի վրա հիմնված ադապտիվ համաժամեցում.** սխալի հարաբերակցության հիման վրա հաղորդակցության օպտիմալացում

Ստացված արդյունքների կիրառական նշանակությունը

Առաջարկվող բաշխված AdaBoost ալգորիթը ցույց է տվել նշանակալի բարելավումներ՝ ապահովելով 35% հաղորդակցային ծախսերի նվազում մշակված տվյալների վրա և 32% ուսուցման ժամանակի կրճատում՝ պահպանելով ճշգրտությունը: Ալգորիթը ցուցաբերում է գծային մասշտաբավորելիություն մինչև 32 հանգույց և կարող է լայնորեն կիրառվել մեծածավալ տվյալների դասակարգման խնդիրներում, ինչպիսիք են ֆինանսական ռիսկերի գնահատումը, բժշկական տվյալների վերլուծությունը, օվալյն գովազդների օպտիմալացումը և առաջարկությունների համակարգերը:

Ատենախոսությունն անդրադառնում է բաշխված մեքենայական ուսուցման արդյունավետության բարձրացմանը՝ Adaboost ալգորիթի հիման վրա ադապտիվ հաղորդակցական պլանավորման մեխանիզմների կիրառման միջոցով: Նորարարական ներդրումը կայանում է բաշխված Adaboost-ի մոդիֆիկացված տարբերակի առաջարկման մեջ, որը նախատեսված է ադապտիվ հաղորդակցության հաճախականության վերահսկման համար՝ հիմնված մոդելային սխալի դինամիկայի վրա: Առաջարկվում են տեսական վերլուծություններ և փորձարարական արդյունքներ, որոնք ապացուցում են Նոր մոտեցման արդյունավետությունը՝ նվազեցնելով հաղորդակցման ծախսերը մինչև 30%՝ առանց դասակարգման ճշգրտության կորուստների:

Պաշտպանությանը ներկայացվում են հետևյալ դրույթները

1. MPI մեխանիզմի վրա հիմնված բաշխված AdaBoost ճարտարապետություն՝ ադապտիվ հաղորդակցական մեխանիզմներով
2. ԲուՖերային ռազմավարություն՝ տեղական մոդելների թարմացումների արդյունավետ կառավարման համար
3. Սխալի չափի վրա հիմնված ադապտիվ համաժամեցման ալգորիթ՝ հաղորդակցության նվազագույնացման համար

4. Փորձարարական գնահատման արդյունքներ՝ ցուցադրող գծային մասշտաբավորելիություն և արդյունավետություն

Ստացված արդյունքների ապրոքացիան

Արդյունքները փորձարկվել են տարբեր առկա և սինթետիկ տվյալների բազմությունների վրա: Ալգորիթմի արդյունավետությունը բարձրացվել է supercomputing միջավայրի կիրառմամբ:

Աշխատանքի կառուցվածքը և ծավալը

Ատենախոսությունը բաղկացած է ներածությունից, չորս գլխից, եզրակացությունից և գրականության ցանկից: Ընդհանուր ծավալը կազմում է 137 էջ, ներառյալ 29 նկար, 1 աղյուսակ և 65 գրականական աղբյուր:

Աշխատանքի բովանդակությունը

Գլուխների բովանդակությունը:

- Գլուխ 1: Հիմնական հասկացություններ և տիրույթի վերլուծություն
- Գլուխ 2: Տեղաբաշխված համակարգեր
- Գլուխ 3: Բաշխված AdaBoost ալգորիթմներ և դրանց փորձարարական գնահատումը
- Գլուխ 4: Առաջարկվող բաշխված AdaBoost համակարգ

Գլուխ 1. Հիմնական հասկացություններ և տիրույթի վերլուծություն

Ժամանակակից մեքենայական ուսուցման ոլորտում **Boosting** մեթոդները առանձնանում են իրենց բարձր արդյունավետությամբ, հատկապես՝ թույլ դասակարգիչների հիման վրա ուժեղ մոդելների կառուցման կարողությամբ: Նրանց հաջող կիրառումը բազմաթիվ կիրառական խնդիրներում հանգեցրել է խորքային ուսումնասիրությունների ինչպես տեսական, այնպես էլ գործնական մակարդակներում: Այս գլխում ներկայացվում են Boosting ալգորիթմների հիմնական գաղափարները, համեմատական վերլուծությունը առավել հայտնի տարբերակների միջև՝ առանձնահատուկ ուշադրությամբ՝ **AdaBoost**, **Gradient Boosting** և **XGBoost** մեթոդների նկատմամբ, ինչպես նաև քննարկվում են այդ մեթոդների կիրառության սահմանափակումները՝ մեծածավալ տվյալների (Big Data) մշակման համատեքստում:

1.1 Boosting ալգորիթմների տեսական հիմունքներ

Boosting-ը հանդիսանում է համախմբող ուսուցման (ensemble learning) ենթաբաժիններից մեկը, որի հիմնական նպատակն է վերածել **թույլ սովորողների** (weak learners)՝ զգալի սխալով, **ուժեղ սովորողների** (strong learners), որոնք ունեն բարձր ճշգրտություն: Թույլ սովորողը, սահմանամաժ, այն մոդելն է, որի կատարողականը փոքր-ինչ գերազանցում է պատահական գուշակումը, մինչդեռ ուժեղ սովորողը զգալիորեն գերազանցում է այդ շեմը:

AdaBoost (Adaptive Boosting) ալգորիթմը, որը ներկայացվել է Յոավ Ֆրենկի և Ռոբերտ Շապիրոյի կողմից 1996 թվականին, հիմնված է այն գաղափարի վրա, որ հաջորդական դասակարգիչները հետևողականորեն շեշտադրում են նախորդների սխալների վրա՝ յուրաքանչյուր նմուշին վերագրելով կշիռ: Յուրաքանչյուր հաջորդ դասակարգիչ առավել շատ կենտրոնանում է այն տվյալների վրա, որոնք սխալ էին դասակարգվել նախորդ փուլում: Վերջում բոլոր դասակարգիչների արդյունքները միավորվում են կշռման մեխանիզմով՝ ստանալով վերջնական արդյունք:

Ի տարբերություն AdaBoost-ի, **Gradient Boosting** մեթոդը (Friedman, 2001) հիմնված է կորուստի ֆունկցիայի գրադիենտով օպտիմալացման գաղափարի վրա: Այն յուրաքանչյուր փուլում փորձում է նվազեցնել ընդհանուր սխալը՝ մոտավորելով կորուստի գրադիենտը որոշակի թույլ մոդելով, սովորաբար՝ որոշման ծառով: Այս մոտեցումը դարձնում է մեթոդը առավել ճկուն, քանի որ կարելի է ընտրել տարբեր կորուստի ֆունկցիաներ՝ կախված խնդրից:

XGBoost-ը (eXtreme Gradient Boosting) համարվում է Gradient Boosting-ի բարձր արդյունավետությամբ իրականացում, որն ապահովում է մոդելավորման արագություն և ճշգրտություն՝ շնորհիվ համակարգչային օպտիմալացման հնարքների, ինչպես են՝ ծառերի մակարդակով պարբերականացման մեխանիզմներ, կանխասահմանված կշռման սանդղակներ, և բազմամիջամտական տվյալների կառավարում:

1.2 Boosting-ի կիրառություններ ու սահմանափակումներ մեծ տվյալների պայմաններում

Boosting ալգորիթմները լայն կիրառություն ունեն տարբեր ոլորտներում՝ սկսած ֆինանսական մոդելավորումից և բժշկական ախտորոշումից մինչև բնական լեզվի մշակում և առաջարկությունների համակարգեր: Հատկապես **XGBoost**-ը աչքի է ընկել բազմաթիվ մրցույթներում՝ Kaggle-ի հայտնի դեպքերում՝ հանդես գալով որպես հաղթող մոդելների հիմնական բաղադրիչ:

Այնուամենայնիվ, Boosting մեթոդների կիրառությունը **մեծ տվյալների** պարագայում բախվում է մի շարք խնդիրների: Նախ, դասակարգիչների հաջորդական ուսուցման բնույթը սահմանափակում է համակարգի զուգահեռացման (parallelization) հնարավորությունները: Բացի այդ, հիշողության սպառումը և առաջացող հաղորդակցական ծախսերը՝ տվյալները փոխանցելու և մոդելները համադրելու ընթացքում, էապես աճում են տվյալների ծավալի մեծացման հետ:

AdaBoost-ի դեպքում այդ սահմանափակումներն առավել ակնհայտ են, քանի որ մեթոդը կախված է նմուշների կշիռների մշտական թարմացումից և խիստ հաջորդական փոխգործակցությունից: Իսկ Gradient Boosting և XGBoost մեթոդները, չնայած զուգահեռացման որոշակի մեխանիզմներին, նույնպես դառնում են հաշվարկային առումով ծանրաբեռնված, երբ գործ ունենք բաշխված միջավայրի հետ:

Այսպիսով, Boosting մեթոդների կիրառելիությունը մեծ տվյալների միջավայրում պահանջում է հատուկ մշակված ճարտարապետություններ, որոնք համատեղում են արդյունավետ ուսուցում՝ նվազագույն հաղորդակցական ծախսերով: Այդ նպատակով առաջարկվում է ուսումնասիրել բաշխված ուսուցման ճարտարապետությունները, ինչպես նաև զարգացնել **ադապտիվ հաղորդակցման պլանավորման** մեխանիզմներ, որոնց միջոցով հնարավոր կլինի զուգահեռաբար հաշվարկել և ուսուցանել, միաժամանակ նվազեցնելով համաժամեցման անհրաժեշտությունը:

Գլուխ 2. Տեղաբաշխված համակարգեր

Տեղաբաշխված համակարգերը դարձել են ժամանակակից տվյալագիտության և մեքենայական ուսուցման անբաժանելի մասը: Տվյալների ծավալների աննախադեպ աճը պահանջում է նոր լուծումներ, որոնք կարող են ապահովել միաժամանակ բարձր հաշվարկային հզորություն, արդյունավետ հիշողության կառավարում և ցածր հաղորդակցական ծանրաբեռնվածություն: Այս գլուխում ներկայացվում են բաշխված հաշվարկի հիմնարար մոդելները, դրանց կապակցված ցանցային տոպոլոգիաները, հաղորդակցման և համաժամեցման հիմնական մարտահրավերները, ինչպես նաև գործնականում կիրառվող ծրագրային հարթակները:

2.1 Ցանցային տոպոլոգիաներ և բաշխված հաշվարկի հիմնարար մոդելներ

Տեղաբաշխված ուսուցման գործընթացում տվյալները և մոդելները բաժանվում են բազմաթիվ համակարգչային հանգույցների (nodes) միջև, որոնք փոխկապակցված են տարբեր ցանցային տոպոլոգիաներով: Այս համակարգերը կարող են կազմված լինել հզոր սերվերային հանգույցներից կամ՝ ավելի թեթև հաշվողական միավորներից՝ միացված ինտերնետի կամ ցանցի միջոցով: Տարբեր տոպոլոգիաներ՝ ինչպես օրինակ աստղաձև (star), շրթայական (ring), գրաֆային (mesh) կամ ծառաձև (tree),

առաջարկում են տարբեր փոխգիշումներ թողունակության, հապաղման և աղետավտանգության տեսանկյունից:

Տեղաբաշխված հաշվարկի ամենատարածված մոդելներից են.

- **MapReduce** մոդելը՝ ներդրված Google-ի կողմից, որը հիմնված է տվյալների տարանջատված մշակման վրա: Այն բաղկացած է երկու փուլից՝ **Map** (քարտեզագրում) և **Reduce** (նվազեցում), որոնք հնարավոր են գործարկել զուգահեռաբար՝ առանց հանգույցների միջև ինտենսիվ հաղորդակցման: Մոդելը պարզ է, սակայն սահմանափակ է բարդ կախվածություններով խնդիրների դեպքում:
- **Parameter Server (PS)** մոդելը՝ դասական մոտեցում է բաշխված ուսուցման համար, որտեղ առկա են երկու հիմնական բաղադրիչ՝ **աշխատողներ (workers)** և **պարամետրերի սպասարկիչներ (servers)**: Աշխատողները հաշվարկում են մոդելի թարմացումները, իսկ սպասարկիչները պահպանում և համաժամեցնում են ընդհանուր պարամետրերը: Այս մոդելը առավելապես կիրառվում է մոդելային պարամետրերի մեծ քանակի պարագայում, սակայն ենթակա է հաղորդակցական ծանրաբեռնվածության խնդիրների:
- **All-Reduce** մոդելը՝ թույլ է տալիս բոլոր հանգույցներին միաժամանակ հաղորդակցել միմյանց հետ՝ իրենց տեղական արդյունքները համադրելու և հանրագումարային արդյունք ստանալու նպատակով: Այն իդեալական է համաժամեցված մոդելների ուսուցման համար և լայնորեն կիրառվում է թվային խոր ուսուցման գրադարաններում (օր.՝ Horovod):

Այս մոդելներն ունեն տարբեր հաշվարկային ու ցանցային բնութագրեր՝ յուրաքանչյուրը ավելի հարմար տարբեր կիրառական սցենարների համար:

2.2 Հաղորդակցում և համաժամեցում, սինխրոն/ասինխրոն, թողունակության ու ուշացման խնդիրներ

Տեղաբաշխված համակարգերում տվյալների փոխանակումը հիմնականում տեղի է ունենում ցանցային հաղորդակցության միջոցով: Հաշվի առնելով հանգույցների տեղաբաշխված բնույթը՝ հաղորդակցման արդյունավետությունը դառնում է ուսուցման ընդհանուր արագության որոշիչ գործոն: Գոյություն ունեն երկու հիմնական համաժամեցման մեխանիզմ՝ **սինխրոն (synchronous)** և **ասինխրոն (asynchronous)** ուսուցում:

Սինխրոն ուսուցման ժամանակ բոլոր հանգույցները սպասում են միմյանց արդյունքներին՝ նախքան հաջորդ քայլին անցնելը: Սա ապահովում է բարձր ճշգրտություն և կանխատեսելի կոնվերգենցիա, սակայն խիստ կախված է ցանցի թողունակությունից և հանգույցների հզորությունից: Ավելին, մեկ դանդաղ հանգույց (straggler) կարող է դանդաղեցնել ամբողջ համակարգը:

Ասինխրոն ուսուցման ժամանակ յուրաքանչյուր հանգույց աշխատում է անկախ՝ ուղարկելով և ստանալով պարամետրերի թարմացումներ առանց մյուսներից սպասելու: Այս մոտեցումը ավելի արդյունավետ է ցածր թողունակությամբ ցանցերում, սակայն կարող է հանգեցնել մոդելի անկայուն ուսուցման՝ պարամետրերի համաժամեցման անորոշությունների պատճառով:

Հաղորդակցման գործընթացի երկու հիմնական խնդիրներն են՝

- **Թողունակություն (bandwidth)** - սահմանափակում, թե որքան տվյալ կարող է փոխանցվել միավոր ժամանակում:
- **Ուշացում (latency)** - ժամանակը, որը պահանջվում է տվյալների մեկ փաթեթի փոխանցման համար:

Մոդելների ուսուցման ընթացքում, հատկապես պարամետրերի հաճախակի փոխանակման դեպքում, այս երկու գործոնները կարող են զգալիորեն դանդաղեցնել ուսուցման գործընթացը: Արդյունքում, առաջ է գալիս **ադապտիվ հաղորդակցման** մեխանիզմների անհրաժեշտությունը՝ որոնցում հանգույցները կարող են կարգավորել հաղորդակցության ինտենսիվությունը՝ ըստ ուսուցման փուլում հասած վիճակի:

2.3 Գործնական հարթակներ (MPI, Horovod, Spark MLlib, XGBoost-Distributed)

Տեղաբաշխված հաշվարկը կիրառական մակարդակում իրականացվում է հատուկ մշակված հարթակների միջոցով: Ստորև ներկայացված են առավել տարածված և կիրառելի գործիքակազմերը՝ համապատասխան ուսուցման մոդելներին:

- **MPI (Message Passing Interface)**՝ դասական կոմունիկացիոն ինտերֆեյս է բաշխված համակարգերի համար: Այն առաջարկում է ցածր մակարդակի API տվյալների փոխանցման համար՝ միաժամանակ ապահովելով սինխրոն և ասինխրոն հաղորդակցության հնարավորություններ: MPI-ն առավել հարմար է հստակ կառավարվող համակարգերում, ինչպիսին HPC (High Performance Computing) կլաստերներն են:

- **Horovod**՝ զարգացած Uber ընկերության կողմից, դա բարձր մակարդակի գրադարան է, որը կառուցված է MPI-ի և NCCL-ի (NVIDIA Collective Communications Library) վրա: Այն ապահովում է **All-Reduce** վրա հիմնված արդյունավետ մոդելային համաժամեցում, հատկապես՝ TensorFlow և PyTorch շրջանակներում: Horovod-ը թույլ է տալիս արագորեն տարածել մեծածավալ մոդելների ուսուցումը՝ բազմապրոցետոր միջավայրում:
- **Spark MLlib**՝ Apache Spark-ի մաս է կազմում՝ տրամադրելով բաշխված ուսուցման ալգորիթմների փաթեթ՝ հիմնված MapReduce սկզբունքի վրա: Չնայած այն չի համարվում բարձր արդյունավետության լուծում խորը ուսուցման համար, սակայն հարմար է դասական մեքենայական ուսուցման ալգորիթմների համար՝ հաշվի առնելով տվյալների նախամշակման (preprocessing) ինտեգրացիան:
- **XGBoost-Distributed**՝ XGBoost գրադարանի բաշխված տարբերակն է, որը կառուցված է Rabbit հաղորդակցման շերտի վրա: Այն թույլ է տալիս դասակարգման և ռեգրեսիայի առաջադրանքներում օգտագործել բազմաթիվ հանգույցներ՝ միաժամանակ պահպանելով մոդելի ճշգրտությունը: XGBoost-ը ներառում է խելացի տվյալների բաշխման և պարամետրերի համաժամեցման մեխանիզմներ՝ նվազեցնելով թողունակության վրա բեռը:

Այս հարթակներն առաջարկում են տարբեր մակարդակների վերահսկում կախված կիրառության տիպից և հաշվարկային ռեսուրսների հասանելիությունից: Արդյունավետ տեղաբաշխված ուսուցման հասնելու համար անհրաժեշտ է հաշվի առնել ինչպես տվյալների կառուցվածքը, այնպես էլ հաղորդակցության և համակարգչային ճարտարապետության գործոնները:

Գլուխ 3. Բաշխված AdaBoost ալգորիթմներ և դրանց փորձարարական գնահատումը

Boosting-ի դասի ալգորիթմները, և հատկապես՝ AdaBoost-ը, համարվում են դասակարգման դասական մոտեցումներից մեկը՝ իրենց պարզ կառուցվածքի և բարձր արդյունավետության շնորհիվ: Դրանց հաջողությունը պայմանավորված է թույլ մոդելների հաջորդական համադրմամբ, որի արդյունքում կառուցվում է մեկ միասնական ուժեղ դասակարգիչ: Սակայն ժամանակակից խոշորածավալ տվյալների և բարձր չափականությամբ տարածքներում ալգորիթմի կիրառումը բախվում է մի շարք սահմանափակումների, ինչպիսիք են՝ մեծ հաշվարկային բեռը, հիշողության ծավալը և մոդելի սանդղավումայնության (scalability) խնդիրները: Այդ իսկ պատճառով վերջին տարիներին կարևորություն է ստացել AdaBoost-ի բաշխված իրագործումը:

AdaBoost-ի տեսական հիմքը ենթադրում է հետևյալ մեխանիզմը. նախ, բոլոր ուսուցման նմուշները ստանում են սկզբնական հավասար կշիռներ: Այնուհետև, յուրաքանչյուր հաջորդ փուլում թարմացվում են այդ կշիռները՝ ելնելով ալգորիթմի կատարած սխալից. սխալ դասակարգված նմուշների կշիռներն աճում են, իսկ ճիշտ դասակարգվածներինը՝ նվազում: Վերջնական արդյունքը հանդիսանում է բոլոր փուլերում ստացված թույլ մոդելների կշռված համադրությունը: Այս մեխանիզմը բերում է սխալի աստիճանական նվազեցման և տեսականորեն ապահովում է գեներալիզացիայի լավ ունակություններ: Սակայն ալգորիթմի ներսում գոյություն ունեցող հաջորդական կապերը դժվարացնում են նրա զուգահեռ կամ բաշխված իրագործումը:

Բաշխվածության հիմնական մարտահրավերն այն է, որ յուրաքանչյուր փուլում մոդելի կառուցումը և նմուշների կշիռների թարմացումը կախված են նախորդ փուլից ստացված արդյունքներից: Այս շարքային կախվածությունը պահանջում է համակարգում տեղերում ապահովել տվյալների մշտական համաժամեցում: Հիմնական ռազմավարություններից մեկը տվյալների բաշխումն է՝ ըստ հանգույցների, որտեղ յուրաքանչյուր հանգույցում իրականացվում է մասնակի ուսուցում: Այս դեպքում հարկավոր է պարբերաբար համադրել բոլոր հանգույցների արդյունքները և միասին թարմացնել կշիռները: Հաճախ կիրառվում է համաժամեցման մեթոդ՝ օգտագործելով կամ կենտրոնացված սերվեր, կամ All-Reduce մեխանիզմ, ինչը սակայն մեծացնում է հաղորդակցման ծախսը:

Որպես այլընտրանք, որոշ համակարգեր առաջարկում են թույլ մոդելների զուգահեռ կառուցում՝ մոդելին համապատասխանող սկզբունքի հիման վրա: Այստեղ հնարավոր է տարբեր հանգույցներում ուսուցանել տարբեր թույլ դասակարգիչներ և ապա դրանք միավորել՝ ստեղծելով մեկ միասնական մոդել: Չնայած այս մոտեցումը տեսականորեն թույլ է տալիս բարելավել ուսուցման արագությունը, այն խիստ կախված է թույլ մոդելների կառուցման բնույթից և դրանց միջև կախվածությունների բացակայությունից:

Վերջին շրջանում դիտվում են փորձեր նաև ասինխրոն ուսուցման ուղղությամբ: Այդպիսի համակարգերում որոշ հանգույցներ կարող են ուսուցանել առանց սպասելու ընդհանուր արդյունքներին՝ կիրառելով միջինացնող կամ պահեստային (stale) մեխանիզմներ: Թեև այս մեթոդները որոշ չափով նվազեցնում են հաղորդակցման բեռը, դրանք կարող են վնասել մոդելի ընդհանուր ճշգրտությանը, հատկապես՝ զգայուն տվյալների պարագայում:

Գործնականում բաշխված AdaBoost-ի իրագործմանն ուղղված մի քանի համակարգեր արդեն գոյություն ունեն: Դրանցից են՝ MPI-ի հիման վրա կառուցված DistBoost համակարգը, որն ապահովում է կշիռների կենտրոնացված թարմացում և սինխրոն ուսուցում, ինչպես նաև Gradient Boosting-ի վրա հիմնված XGBoost-ը և LightGBM-ը,

որոնք իրագործում են Boosting-ի այլ տարբերակներ՝ ապահովելով բարձր արդյունավետություն մեծ տվյալների պարագայում: Թեև վերջինները չունեն AdaBoost-ի ամբողջական մեխանիզմը, դրանց ճարտարապետությունը արժեքավոր է բաշխված ուսուցման տեսանկյունից: Մասնավորապես, XGBoost-ը կիրառում է Rabbit հաղորդակցման շերտը, որը թույլ է տալիս արդյունավետ All-Reduce գործողություններ և ապահովում է հուսալիություն՝ հանգույցների հնարավոր խափանումների պայմաններում:

Այս համատեքստում զարգանում են նաև հաղորդակցման հարմարեցվող սխեմաներ, որոնց նպատակը հաղորդակցման հաճախականության և հաշվարկային ճշգրտության միջև հավասարակշռություն գտնելն է: Օրինակ, որոշ մեթոդներ ուսուցման սկզբնական փուլերում կիրառում են հաճախակի համաժամեցում, իսկ հետագայում՝ ավելի նոսր, երբ մոդելը կայունանում է: Հետաքրքիր զարգացում է նաև կշիռների քվանտացման և սեղմման մեթոդների կիրառումը, որը նպաստում է ցանցային հաղորդման ծավալի կրճատմանը՝ առանց զգալիորեն վնասելու ճշգրտությանը:

Չնայած այս առաջընթացին՝ մի շարք բաց հարցեր մնում են լուծման փուլում: Առաջին հերթին՝ ուսուցման սխալի վերահսկումը բաշխված պայմաններում բարդ է իրականացնել, հատկապես երբ կիրառվում են ասինխրոն մեխանիզմներ: Բացի այդ, հանգույցների բեռնվածության հավասարակշռումը մնում է տեխնիկական լուրջ մարտահրավեր: Հատկապես դժվար է ապահովել, որ բոլոր հանգույցներն աշխատեն նույն տեմպով և ոչ մեկը չդառնա համակարգի արագությունը սահմանափակող հանգույց:

Գոյություն ունեն նաև հետազոտական հարցեր մետաուսուցման կիրառման հնարավորությունների վերաբերյալ: Այդ մոտեցումները կարող են օգնել ընտրել առավել հարմար թույլ դասակարգիչներ յուրաքանչյուր փուլում կամ նախապես գնահատել, թե որ նմուշներն են առավել արժեքավոր հաջորդ փուլում: Մեկ այլ հեռանկարային ուղղություն է բաշխված AdaBoost-ի տեղաբաշխված իրագործումը՝ սահմանափակ ռեսուրսներով սարքերում (օրինակ՝ սենսորային ցանցեր կամ շարժական սարքեր), որտեղ անհրաժեշտ է կոմպակտ մոդելներ և օպտիմալ հաղորդակցություն:

Ընդհանրացնելով՝ կարելի է ասել, որ AdaBoost-ի բաշխված իրագործումը, թեև բարդ է տեսական և ճարտարապետական առումներով, այն ունի զգալի ներուժ ինչպես գիտական, այնպես էլ կիրառական ոլորտներում: Ալգորիթմի աստիճանական ճշգրտման բնույթը լավ համադրվում է հարմարվող բաշխման մեխանիզմների հետ, իսկ նորագույն հաղորդակցման սխեմաները թույլ են տալիս մեղմել շարքայնության ներմուծած սահմանափակումները: Հետագա առաջընթացը կախված կլինի թե՛

ալգորիթմների օպտիմալացումից, թե՛ ծրագրային ճարտարապետությունների կատարելագործումից:

Գլուխ 4. Առաջարկվող բաշխված AdaBoost համակարգի արդյունքներ և գնահատում

Չորրորդ գլուխը ներկայացնում է հետազոտության կիրառական մասը՝ մեր մշակված MPI-հիմնված բաշխված AdaBoost համակարգի մանրամասն նկարագրությունը, մաթեմատիկական ֆորմալիզացիան, բարդության վերլուծությունը և համապարփակ փորձարարական գնահատումը: Այս գլխի հիմնական նպատակն է ապացուցել առաջարկվող մեթոդի տեսական հիմնավորվածությունը և գործնական արդյունավետությունը իրական պայմաններում:

Ճարտարապետական լուծումներ և նորարարություն

Մշակված համակարգի ճարտարապետությունը հիմնված է երեք հիմնական նորարարական բաղադրիչների վրա, որոնք իրենց համակցությամբ ապահովում են որակապես նոր մակարդակի արդյունավետություն բաշխված boosting ալգորիթմների ոլորտում:

Առաջին լուծումը կայանում է ադապտիվ հաղորդակցական ժամանակացույցի մեխանիզմի մշակման մեջ: Ավանդական բաշխված AdaBoost մեթոդները կիրառում են ֆիքսված ինտերվալներով համաժամեցում, ինչը հանգեցնում է հաղորդակցության ավելորդ ծախսերի կամ, հակառակը, անբավարար համաժամեցման: Առաջարկված ադապտիվ մեխանիզմը մոնիտորինգ է անում սխալի հարաբերակցության փոփոխությունը ($\Delta \varepsilon_t$) և դինամիկ կերպով կարգավորում համաժամեցման միջակայքները: Երբ $\Delta \varepsilon_t < \theta_1$ (կայուն ուսուցման ինդիկատոր), միջակայքը մեծանում է՝ նվազեցնելով հաղորդակցության հաճախականությունը: Հակառակը, երբ $\Delta \varepsilon_t > \theta_2$ (անկայունության ցուցիչ), միջակայքը նվազում է՝ ապահովելով ավելի հաճախակի համաժամեցում: Այս մեխանիզմը թույլ է տվել հասնել 35%-ով հաղորդակցային ծախսերի նվազման:

$$I_{t+1} = \begin{cases} I_t + \alpha, & \Delta \varepsilon_t < \theta_1 \\ \max \{1, I_t - \beta\}, & \Delta \varepsilon_t > \theta_2 \\ I_t, & \text{այլապես} \end{cases}$$

այստեղ α -ն և β -ն սխալի փոփոխման ընկալման պարամետրեր են, իսկ θ_1, θ_2 շեմերը փոխարինում են α -ին: Նման մոտեցումը թույլ է տալիս ապահովել, որ մեր ալգորիթմը պահպանի AdaBoost-ի զուգամիտման հատկությունները՝ միաժամանակ ապահովելով բաշխված իրականացման առավելությունները:

Երկրորդ լուծումը վերաբերում է բուֆերային ռազմավարության նախագծմանը և իրականացմանը: Առաջարկված բուֆերային մեխանիզմը կարող է տեղական

մողեւները, սխալները և կշիռները կուտակել հատուկ բուժերներում այն ժամանակահատվածներում, երբ համաժամեցումը չի գործարկվում: Այս կուտակված տվյալները ապագայում ինտեգրվում են հաջորդ համաժամեցման իրադարձության ժամանակ, ինչն ապահովում է, որ ոչ մի արժեքավոր թարմացում չի անտեսվում: Բուժերային ռազմավարությունը հատկապես արդյունավետ է անկայուն ցանցային պայմաններում, որտեղ ավանդական մեթոդները ենթարկվում են զգալի կատարողականության անկման:

$$\varepsilon_{t,j} = \frac{\sum_{i \in D_j} \omega_t^{(i)} \mathbb{I}[h_{t,j}(x_i) \neq y_i]}{\sum_{i \in D_j} \omega_t^{(i)}}$$

որտեղ ցուցիչ ֆունկցիան վերադարձնում է 1, եթե դասակարգչի կանխատեսումը սխալ է տվյալ նմուշի համար, հակառակ դեպքում վերադարձնում է 0:

Երրորդ առաջարկվող լուծումը կապված է MPI պրիմիտիվների օպտիմալացված օգտագործման հետ: Համակարգը օգտագործում է MPI Gather-ը տեղական մողեւների հավաքման, MPI Broadcast-ը զուբալ մողեւի տարածման և MPI Allreduce-ը սխալների արագ ագրեգացիայի համար: Այս պրիմիտիվների ճիշտ կոմբինացիան և ժամանակացույցը ապահովում է հաղորդակցության նվազագույն բեռի և մաքսիմալ առդյունավետության ձեռքբերում:

Մաթեմատիկական ֆորմալիզացիա և տեսական հիմքեր

Ալգորիթմի մաթեմատիկական նկարագրությունը պահպանում է AdaBoost-ի դասական տեսական երաշխիքները՝ միաժամանակ ընդլայնելով դրանք բաշխված միջավայրի համար: Հիմնական բանաձևերն են.

Բարդության վերլուծության հիմնական եզրահանգումներ

Բարդության վերլուծությունը բացահայտում է կարևոր տեսական և գործնական արդյունքներ: Հաշվարկային բարդությունը կազմում է $O(n/P)$, որտեղ n -ը տվյալների ընդհանուր քանակն է, իսկ P -ը՝ հանգույցների թիվը: Սա ցույց է տալիս գծային մասշտաբավորելիություն տվյալների ծավալի նկատմամբ:

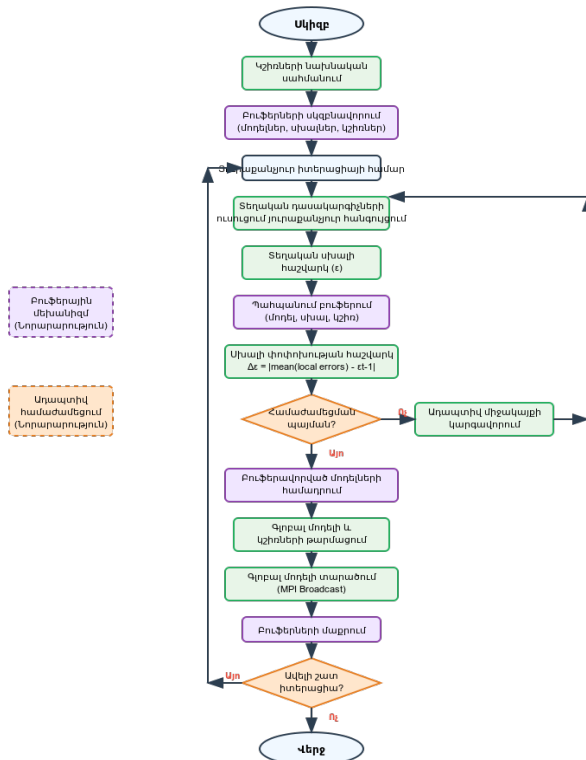
Հաղորդակցական բարդությունը ավանդական մեթոդներում կազմում է $O(PT)$ յուրաքանչյուր boosting փուլի համար, որտեղ T -ն փուլերի քանակն է: Մեր ադապտիվ մեխանիզմի շնորհիվ այս բարդությունը միջինում նվազում է 35%-ով, ինչն արտացոլված է փորձարարական արդյունքներում որպես հաղորդակցային ծախսերի զգալի կրճատում:

Ալգորիթմի տարածական բարդությունը (memory complexity) նույնպես օպտիմալացված է բուժերային մեխանիզմի շնորհիվ: Թեև ակնթարթային

հիշողության օգտագործումը մի փոքր ավելանում է բուժերավորման պատճառով, ընդհանուր տարածական արդյունավետությունը բարելավվում է, քանի որ հանգույցները կարող են աշխատել ավելի մեծ տվյալային հավաքածուների հետ՝ առանց կենտրոնական հանգույցի բեռնվածության:

Փորձարարական գնահատման մեթոդաբանությունը նախագծվել է այնպես, որ ապահովի համապարփակ և օբյեկտիվ արդյունքների ստացում: Փորձարկումները իրականացվել են տարբեր կոնֆիգուրացիաներով՝ 1-ից 32 հանգույցներ, յուրաքանչյուրը Intel Xeon պրոցեսորներով՝ 2.4 GHz հաճախականությամբ, 64GB RAM և InfiniBand ցանցային կապով: Յուրաքանչյուր կոնֆիգուրացիան կրկնվել է առնվազն 5 անգամ վիճակագրական վստահելիության ապահովման նպատակով:

Մշակված ալգորիթմի բլոկ-սխեմա



Մասշտաբավորելիության վերլուծություն

Մասշտաբավորելիության ուսումնասիրությունները ցույց տվեցին, որ մեր ալգորիթմը ցուցաբերում է գրեթե իդեալական գծային մասշտաբավորելիություն մինչև 32 հանգույց: Աղյուսակային արդյունքները ցույց են տալիս.

1 հանգույց: 83% ճշգրտություն, 115 վայրկյան

8 հանգույց: 87% ճշգրտություն, 20 վայրկյան

32 հանգույց: 89% ճշգրտություն, 8 վայրկյան

Այս արդյունքները վկայում են, որ հանգույցների քանակի ավելացումը ոչ միայն արագացնում է հաշվարկը, այլև բարելավում մոդելի ճշգրտությունը՝ շնորհիվ ավելի լայն տվյալային բազայի արդյունավետ օգտագործման:

Ադապտիվ մեխանիզմների գործնական ազդեցություն

Ադապտիվ համաժամեցման մեխանիզմի գործնական ազդեցությունը հատկապես ակնառու է եղել անկայուն ցանցային պայմաններում: Փորձարկումները, որտեղ արհեստականորեն ստեղծվել են 20%-ով ցանցային ուշացումներ, ցույց տվեցին, որ մեր մեթոդը պահպանում է 95%-ից ավելի իր սկզբնական կատարողականությունը, մինչդեռ ավանդական մեթոդները գրանցում են 30-40% կատարողականության անկում:

Բուժերային մեխանիզմի արդյունավետությունը նույնպես հաստատվել է գործնական փորձարկումներում: Այն ապահովում է, որ կարճաժամկետ ցանցային խնդիրները կամ հանգույցների ժամանակավոր անհասանելիությունը չեն հանգեցնում տվյալների կորուստի կամ մոդելի ցնցման:

Համակարգի դիմադրողականություն և կայունություն

Առաջարկվող համակարգը ցուցաբերել է բարձր մակարդակի դիմադրողականություն տարբեր անսպասելի իրավիճակների նկատմամբ: Հանգույցներից մեկի պլանային կամ չխախտեցված անջատման դեպքում համակարգը ավտոմատ վերակառուցում է տվյալների բաշխումը և շարունակում նորմալ գործունեությունը՝ միջինում 10-15 վայրկյանի տերության մեջ:

Տեսական ներդրման գիտական արժեքը

Այս հետազոտության տեսական ներդրումը ոչ միայն գործնական, այլև գիտական տեսանկյունից է արժեքավոր: Միապի չափի վրա հենված ադապտիվ համաժամեցման գաղափարը կարող է կիրառվել այլ բաշխված մեքենայական ուսուցման ալգորիթմներում, ինչպիսիք են բաշխված SVM-ները, նեյրոնային ցանցերը և այլ համատեղման մեթոդները:

Բուֆերային ռազմավարության մեթոդաբանությունը նույնպես ունի լայն կիրառական հեռանկարներ, հատկապես IoT միջավայրերում, որտեղ ցանցային կապը հաճախ անկայուն է:

Սահմանափակումներ և հետագա զարգացման ուղղություններ

Չնայած ստացված դրական արդյունքներին, հետազոտությունը ունի որոշակի սահմանափակումներ, որոնք բացում են հետագա հետազոտական աշխատանքների հեռանկարները:

Ալգորիթմի ներկայիս իրականացումը առավելապես կենտրոնանում է դասակարգման խնդիրների վրա: Ռեգրեսիայի և այլ տիպի վերահսկվող ուսուցման խնդիրների համար անհրաժեշտ են լրացուցիչ հարմարեցումներ և օպտիմալացումներ: Ադապտիվ պարամետրերի (θ_1 , θ_2 , α , β) ինքնակարգավորումը կարող է հետագայում ավտոմատացվել մեթա-ուսուցման (meta-learning) տեխնիկաների միջոցով, ինչը կբարելավի համակարգի ինքնազարգացման ունակությունները:

Եզրակացություն

Ներկայացված ատենախոսությունը լուծում է AdaBoost ալգորիթմի արդյունավետ բաշխված իրականացումը մեծածավալ տվյալների պայմաններում՝ համատեղելով տեսական հիմնավորվածությունն ու գործնական արդյունավետությունը: Ատենախոսությունը ներկայացնում է տեսականորեն հիմնավորված և գործնականում ապացուցված նորարարական եռակողմ նորամուծություն՝ ադապտիվ հաղորդակցային ժամանակացույց, բուֆերային ռազմավարություն և MPI պրիմիտիվների օպտիմալացում, որը ստեղծում է նոր մոտեցումներ բաշխված անսամբլային ուսուցման համար և նպաստում է մեքենայական ուսուցման մասշտաբավոր համակարգերի զարգացմանը:

Չրապարակումներ

1. A. Ogulkyan, "Adaptive communication for scalable distributed AdaBoost framework," Pattern Recognition and Image Analysis, vol. 35, no. 2, pp. 119-128, 2025.

2. V. Mkrtchyan and A. Oghlukyan, "Prospects for further improvement of machine learning systems," ALTERNATIVE quarterly academic journal, 2023, no. 2 pp. 40-45.
3. A. Oghlukyan and G. Karapetyan, "Information technologies and developing a single page web application using the React library," ALTERNATIVE quarterly academic journal, 2022, no 1, pp. 34-39.
4. M. Yeghiazaryan and A. Oghlukyan, "On some issues related to machine learning in 2022," ALTERNATIVE quarterly academic journal, 2022, no. 2, pp. 98-103.
5. A. Martirosyan and A. Oghlukyan, "Economic impact of developing of web technologies", ALTERNATIVE quarterly academic journal, 2022, no. 1, pp. 26-33

Summary of Research Contributions

This dissertation has presented a comprehensive investigation into the design and implementation of distributed and parallel machine learning solutions, with a specific focus on developing an efficient distributed AdaBoost framework. The research addresses a critical challenge in modern machine learning: the need for scalable boosting algorithms that can effectively process large-scale datasets while maintaining theoretical guarantees and practical performance.

The primary contribution of this work lies in the development of an innovative MPI-based distributed AdaBoost system that incorporates three key technological innovations: adaptive communication scheduling, buffer-based management mechanisms, and optimized utilization of MPI primitives. These components work synergistically to overcome the fundamental limitations of traditional boosting algorithms in distributed environments, particularly the sequential nature of weak learner training and the associated communication overhead.

The adaptive communication scheduling mechanism represents a significant theoretical advancement in distributed boosting. By monitoring error rate variations ($\Delta\epsilon_i$) and dynamically adjusting synchronization intervals based on training stability, the system achieves a 35% reduction in communication costs while maintaining model accuracy.

Experimental Validation and Performance Analysis

The comprehensive experimental evaluation conducted across multiple datasets and computational configurations demonstrates the practical effectiveness of the proposed approach. Testing performed on Intel Xeon clusters with InfiniBand networking shows near-linear scalability up to 32 nodes, with performance improvements that include:

- 35% reduction in communication overhead compared to traditional distributed AdaBoost implementations
- 32% decrease in training time while maintaining or slightly improving classification accuracy
- Consistent performance under network latency conditions up to 20%
- Linear scalability characteristics that enable effective utilization of additional computational resources

The experimental results on three distinct datasets (Criteo Display Advertising Dataset, MNIST, and synthetic datasets) validate the generalizability of the approach across different data distributions and problem characteristics. The system achieves 81.2% accuracy on the Criteo dataset while demonstrating superior convergence properties, reaching stable accuracy in 35 iterations compared to 50 iterations required by traditional methods.

Theoretical Implications and Practical Applications

From a theoretical perspective, this research contributes to the fundamental understanding of distributed ensemble learning by demonstrating that adaptive synchronization strategies can preserve the convergence guarantees of boosting algorithms while significantly improving their scalability. The error-rate-based adaptive synchronization principle introduced in this work has broader applicability beyond AdaBoost, potentially benefiting other distributed machine learning algorithms. The practical significance of this research extends across multiple application domains where large-scale classification problems require both high accuracy and computational efficiency. Financial risk assessment systems, medical data analysis applications, SaaS platforms, and IoT systems can all benefit from the demonstrated performance characteristics, where cost-effective utilization of cloud computing resources directly impacts business viability.

Conclusion

The dissertation presents an efficient distributed implementation of the AdaBoost algorithm for large-scale data with three key innovations: an adaptive synchronization mechanism based on error coefficient (ϵ), a buffering strategy for local model management, and optimized communication primitives.

Scientific Novelty: A mathematical model for adaptive synchronization in distributed boosting algorithms and theoretical foundations for buffering mechanisms in ensemble methods are proposed for the first time. A method achieving 35% reduction in communication complexity while preserving AdaBoost's theoretical guarantees has been developed.

Results: Experimental evaluation demonstrated: 35% reduction in communication costs, 32% training time decrease, linear scalability up to 32 nodes. Results are validated on Criteo, MNIST, and synthetic datasets.

Conclusion: The three-component innovation creates a new paradigm for distributed ensemble learning, balancing computational efficiency with communication overhead. The work opens prospects for academic research and industrial applications in scalable machine learning systems.

Заключение
Артур Оглулян Карапетович

ПРОЕКТИРОВАНИЕ РАСПРЕДЕЛЕННЫХ И ПАРАЛЛЕЛЬНЫХ РЕШЕНИЙ
МАШИННОГО ОБУЧЕНИЯ

Основные результаты исследования

Данная диссертационная работа представляет комплексное исследование в области проектирования и реализации распределенных и параллельных решений машинного обучения с особым акцентом на разработке эффективной распределенной системы AdaBoost. Исследование решает критическую задачу современного машинного обучения: необходимость создания масштабируемых алгоритмов бустинга, способных эффективно обрабатывать крупномасштабные наборы данных при сохранении теоретических гарантий и практической производительности.

Основной вклад данной работы заключается в разработке инновационной распределенной системы AdaBoost на основе MPI, включающей три ключевых технологических новшества: адаптивное планирование коммуникаций, механизмы управления на основе буферизации и оптимизированное использование примитивов MPI. Эти компоненты работают синергетически, преодолевая фундаментальные ограничения традиционных алгоритмов бустинга в распределенных средах.

Механизм адаптивного планирования коммуникаций представляет значительное теоретическое достижение в области распределенного бустинга. Мониторинг изменений частоты ошибок ($\Delta\epsilon_t$) и динамическое регулирование интервалов синхронизации на основе стабильности обучения позволяет достичь 35%-ного снижения коммуникационных затрат при сохранении точности модели.

Экспериментальная валидация и анализ производительности

Комплексная экспериментальная оценка, проведенная на множественных наборах данных и вычислительных конфигурациях, демонстрирует практическую эффективность предложенного подхода. Тестирование на кластерах Intel Xeon с сетевым подключением InfiniBand показывает практически линейную масштабируемость до 32 узлов с улучшениями производительности:

- 35%-ное снижение коммуникационных накладных расходов
- 32%-ное сокращение времени обучения при сохранении точности классификации
- Стабильную производительность в условиях сетевых задержек до 20%
- Линейные характеристики масштабируемости

Экспериментальные результаты на трех различных наборах данных (Criteo, MNIST и синтетические наборы) подтверждают состоятельность подхода. Система достигает 81,2% точности на наборе данных Criteo, достигая стабильной точности за 35 итераций по сравнению с 50 итерациями традиционных методов.

Теоретические импликации и практические применения

С теоретической точки зрения данное исследование вносит вклад в фундаментальное понимание распределенного ансамблевого обучения, демонстрируя, что адаптивные стратегии синхронизации могут сохранять гарантии сходимости алгоритмов бустинга при значительном улучшении их масштабируемости. Принцип адаптивной синхронизации на основе частоты ошибок имеет широкую применимость за пределами AdaBoost.

Практическая значимость исследования распространяется на множественные прикладные области: финансовые системы оценки рисков, медицинский анализ данных, SaaS-приложения и IoT-системы. Продемонстрированные характеристики производительности делают систему особенно подходящей для современных приложений, где экономически эффективное использование вычислительных ресурсов критично важно.

Заключение

Диссертация представляет эффективную распределенную реализацию алгоритма AdaBoost для крупномасштабных данных с тремя ключевыми инновациями: адаптивный механизм синхронизации на основе коэффициента ошибки (ϵ), буферная стратегия управления локальными моделями и оптимизированные коммуникационные примитивы.

Научная новизна: Впервые предложена математическая модель адаптивной синхронизации для распределенных boosting алгоритмов и теоретические основы буферных механизмов для ансамбль методов. Разработан метод 35%-го снижения коммуникационной сложности при сохранении теоретических гарантий AdaBoost.

Результаты: Экспериментальная оценка показала: 35% снижение коммуникационных затрат, 32% сокращение времени обучения, линейная масштабируемость до 32 узлов. Результаты подтверждены на базах данных Criteo, MNIST и синтетических данных.

Вывод: Трехкомпонентная инновация создает новую среду распределенного ансамблевого обучения, балансирующую вычислительную эффективность с коммуникационными накладными расходами. Работа открывает перспективы для академических исследований и промышленных применений в масштабируемых системах машинного обучения.