ЛЕКЦИЯ 6
Машиналық оқыту негізінде желілік аномалияларды анықтау

Лекция мақсаты: Желілік аномалия ұғымын, қалыпты желілік мінез-құлық профилін қалыптастыру қағидаларын, машиналық оқыту арқылы аномалияларды анықтау кезеңдерін, қолданылатын белгілер мен алгоритмдерді, модель сапасын бағалау тәсілдерін және олардың кибершабуылдарды анықтаудағы ерекшеліктерін жан-жақты қарастыру.

Желілік аномалия ұғымы

Желілік аномалия — белгілі бір ақпараттық жүйе немесе желі үшін қалыптасқан қалыпты мінез-құлықтан елеулі ауытқу. Мұндай ауытқу трафик көлемінің күрт өсуі, бір құрылғының көптеген мекенжайлармен байланысуы, бұрын қолданылмаған порттың пайда болуы, байланыстар жиілігінің өзгеруі немесе пайдаланушыға тән емес уақытта белсенділіктің байқалуы түрінде көрінуі мүмкін. Аномалияның өзі міндетті түрде кибершабуыл болып саналмайды. Ол бағдарламалық жаңарту, резервтік көшіру, жаңа сервис іске қосу немесе пайдаланушылар санының өзгеруі сияқты заңды процестің нәтижесі де болуы ықтимал. Сондықтан аномалияны анықтау жүйесінің мақсаты — күмәнді ауытқуды бөліп көрсету және оны кейінгі талдауға ұсыну.

Дәстүрлі сигнатуралық жүйелер белгілі шабуыл үлгілерін іздейді. Ал аномалиялық тәсіл алдын ала нақты шабуыл сигнатурасы болмаған жағдайда да қалыпты профильден ерекшеленетін мінез-құлықты байқауға бағытталады. Бұл қасиет жаңа немесе өзгертілген шабуылдарды анықтауда маңызды. Сонымен қатар қалыпты белсенділіктің өзі өте әртүрлі болғандықтан, аномалиялық жүйелерде жалған оң ескертулер мәселесі туындайды.

Қалыпты желілік мінез-құлық профилі

Аномалияны анықтау үшін алдымен қалыпты желілік мінез-құлықтың базалық профилі қалыптастырылады. Профиль белгілі бір уақыт аралығындағы желілік сессиялар, протоколдар, байланыс бағыттары, пакеттер саны, берілген және қабылданған деректер көлемі, сессия ұзақтығы және байланыс жиілігі сияқты сипаттамалар негізінде құрылады. Қалыпты профиль әмбебап болмайды: университет желісіндегі қалыпты трафик өндірістік кәсіпорын немесе банктің желілік трафигінен ерекшеленеді.

Профиль тәуліктің уақытына да тәуелді болуы мүмкін. Мысалы, оқу орнында күндізгі уақытта пайдаланушылар саны жоғары, ал түнде трафик төмен болуы қалыпты жағдай. Егер модель уақыт факторын ескермесе, күндізгі қалыпты жүктемені аномалия деп белгілеуі немесе түнгі күмәнді белсенділікті байқамай қалуы мүмкін. Сондықтан уақыттық және контекстік белгілер желілік аномалияларды анықтауда маңызды.

Желілік деректер және белгілер

Машиналық оқыту алгоритмі желілік трафикті адам сияқты тікелей түсінбейді. Модельге трафик сандық немесе категориялық белгілер жиыны ретінде беріледі. Белгі — зерттелетін желілік объектіні сипаттайтын өлшенетін қасиет. Бір объект жеке пакет, желілік ағын, сессия немесе белгілі уақыт терезесіндегі агрегатталған белсенділік болуы мүмкін.

Белгі	**Мазмұны**	**Аномалияны анықтаудағы маңызы**
Duration	Сессияның ұзақтығы	Өте қысқа немесе әдеттен тыс ұзақ байланыстарды бөлуге көмектеседі
Protocol	TCP, UDP және т.б.	Құрылғы үшін тән емес протоколды байқауға мүмкіндік береді

Source/Destination port	Бастапқы және мақсатты порт	Сирек қолданылатын сервис немесе жаңа байланыс бағытын көрсетеді
Packet count	Пакеттер саны	Белсенділіктің күрт өсуін анықтауға көмектеседі
Byte count	Берілген деректер көлемі	Әдеттен тыс үлкен дерек алмасуды байқауға мүмкіндік береді
Connection rate	Уақыт бірлігіндегі байланыстар саны	Қысқа уақытта көптеген байланыстардың пайда болуын көрсетеді
Failed connections	Сәтсіз байланыстар саны	Қайталанатын күмәнді әрекеттің индикаторы болуы мүмкін

Белгілерді таңдау кезінде олардың қауіпсіздік тұрғысынан ақпараттылығы ескеріледі. Белгілер тым аз болса, модель маңызды заңдылықты көрмеуі мүмкін; шамадан тыс көп және мағынасыз белгілер модельді күрделендіріп, есептеу шығынын арттырады. Сондықтан feature selection және feature engineering кезеңдері модельді құру процесінің маңызды бөлігі.

Деректерді алдын ала өңдеу

Нақты желілік деректер машиналық оқытуға бірден жарамды бола бермейді. Алдымен қайталанған жазбалар, бос мәндер, қате форматтар және қажетсіз белгілер тексеріледі. Категориялық белгілер сандық форматқа түрлендіріледі. Әртүрлі масштабтағы сандық белгілер үшін нормалау немесе стандарттау қолданылуы мүмкін. Мысалы, порт нөмірі мыңдық мәнде, ал белгілі бір коэффициент 0 мен 1 аралығында болуы мүмкін; масштаб айырмашылығы кейбір алгоритмдерге әсер етеді.

Деректерді оқыту және тест жиындарына дұрыс бөлу де өте маңызды. Егер бір шабуыл сессиясының бір-біріне өте ұқсас жазбалары әрі train, әрі test жиынына түссе, модель тест дерегін іс жүзінде алдын ала көргендей болады. Бұл data leakage деп аталады және бағалау нәтижесін жасанды түрде жоғарылатады. Уақыттық желілік деректер үшін өткен кезеңді оқытуға, кейінгі кезеңді тестілеуге пайдалану жиі шынайырақ бағалау береді.

Машиналық оқыту тәсілдері

Әдіс	**Негізгі қағидасы**	**Артықшылығы**	**Кемшілігі**
Random Forest	Көптеген шешім ағаштарының нәтижесін біріктіреді	Тұрақты, күрделі байланыстарды модельдейді, белгілер маңызын бағалауға болады	Белгіленген деректер қажет; үлкен модель ресурс талап етеді
K-Means	Ұқсас объектілерді K кластерге бөледі	Класс белгілерін қажет етпейді, деректер құрылымын зерттеуге ыңғайлы	K мәнін алдын ала таңдау керек; кластер әрқашан қауіп класын білдірмейді
Isolation Forest	Сирек объектілерді аз бөлулер арқылы оқшаулайды	Аномалия детекциясына тікелей бағытталған	Нәтиже шекті мәнге және деректер сипатына тәуелді
Autoencoder	Кіріс деректерін ықшам кодтап, қайта қалпына келтіреді	Күрделі қалыпты үлгілерді үйренуге мүмкіндік береді	Оқыту күрделірек; түсіндіру қиын болуы мүмкін
LSTM	Уақыт бойынша тәуелділіктерді	Оқиғалар реттілігі мен уақыттық трафикке қолайлы	Көп дерек пен есептеу ресурсын қажет етеді

	үйренетін рекурренттік нейрондық желі		

Бақыланатын оқыту

Бақыланатын оқытуда әр мысалға алдын ала дұрыс класс белгісі беріледі. Мысалы, желілік сессиялар «қалыпты» және «шабуыл» деп белгіленеді. Модель осы мысалдар арасындағы заңдылықтарды үйреніп, кейін жаңа сессияның класын болжайды. Бұл тәсілдің маңызды артықшылығы — нәтижені confusion matrix және басқа метрикалар арқылы нақты бағалауға болады.

Мысал. Оқу желісінен алынған тарихи деректерде қалыпты трафик пен бұрын расталған шабуыл оқиғалары бар делік. Random Forest осы деректермен оқытылады. Жаңа желілік сессия келгенде модель оның белгілерін талдап, шабуыл класына жату ықтималдығын береді. Бірақ жаңа шабуыл бұрынғы оқыту деректерінен мүлде өзгеше болса, модель оны дұрыс танымай қалуы мүмкін.

Бақыланбайтын оқыту және аномалия детекциясы

Бақыланбайтын оқытуда деректердің дұрыс класын көрсететін белгі жоқ. Алгоритм деректердің ішкі құрылымын өзі іздейді. K-Means ұқсас желілік сессияларды кластерлерге бөледі. Қалыпты трафиктің негізгі бөлігі ірі кластерлерге жиналып, сирек кездесетін объектілер бөлек орналасуы мүмкін. Бірақ сирек объект автоматты түрде шабуыл емес, сондықтан оның контекстін тексеру қажет.

Isolation Forest аномалияларды анықтау үшін объектіні кездейсоқ бөлулер арқылы оқшаулау идеясын қолданады. Сирек және өзгеше объектілерді оқшаулау үшін әдетте аз қадам қажет болады. Нәтижесінде әр объект үшін аномалия ұпайы есептеліп, белгілі шекті мәннен жоғары объектілер күмәнді деп белгіленеді.

Терең оқыту және уақыттық тәуелділік

Желілік белсенділік көбіне уақыттық реттілікке ие. Бір ғана пакет немесе сессия қалыпты көрінуі мүмкін, бірақ бірнеше минуттағы оқиғалар тізбегі күмәнді мінез-құлықты көрсетуі ықтимал. LSTM тәрізді модельдер алдыңғы уақыт қадамдарының ақпаратын ескеріп, осындай тәуелділіктерді үйренуге мүмкіндік береді. Бұл қасиет желілік трафиктің уақыттық динамикасын зерттеуде пайдалы.

Autoencoder қалыпты трафикпен оқытылған жағдайда қалыпты объектілерді жақсы қалпына келтіруге үйренеді. Қалыпты үлгіден қатты ерекшеленетін жаңа объект үшін reconstruction error жоғары болуы мүмкін. Осы қате аномалия ұпайы ретінде пайдаланылуы ықтимал. Дегенмен шекті мәнді дұрыс таңдау қажет.

Желілік аномалияларды анықтау процесі

Желілік трафик → деректерді жинау → тазалау және қалыпқа келтіру → белгілерді қалыптастыру → Train/Test бөлу → модельді оқыту → аномалия ұпайын немесе класс ықтималдығын есептеу → шекті мән → ескерту → қауіпсіздік маманының тексеруі

Процестің бірінші кезеңінде трафиктен қажетті телеметрия жиналады. Екінші кезеңде деректер ML алгоритміне жарамды форматқа келтіріледі. Үшінші кезеңде қауіпсіздік тұрғысынан маңызды белгілер құрылады. Модель оқыту деректерінен заңдылықты үйренеді, ал тест жиыны оның бұрын көрмеген деректердегі қабілетін бағалайды. Жұмыс режимінде әр жаңа объект үшін класс немесе аномалия ұпайы есептеледі. Соңғы шешім шекті мән мен ұйымның қауіп саясатына тәуелді.

Класс теңгерімсіздігі мәселесі

Нақты желіде қалыпты трафик көлемі шабуыл трафигінен әлдеқайда жоғары болуы мүмкін. Мысалы, 10 000 оқиғаның 9 900-і қалыпты және тек 100-і шабуыл болсын. Егер модель барлық оқиғаны «қалыпты» деп белгілесе, оның Accuracy көрсеткіші 99% болады. Бірақ ол 100 шабуылдың біреуін де таппайды. Сондықтан мұндай жағдайда Accuracy жалғыз өзі модель сапасын сипаттай алмайды.

Класс теңгерімсіздігі кезінде Precision, Recall және F1-score сияқты метрикаларға ерекше назар аударылады. Сонымен қатар train жиынында кластардың таралуын басқару, class weight қолдану немесе деректерді қайта іріктеу тәсілдері пайдаланылуы мүмкін. Бірақ кез келген өзгеріс тест жиынының шынайы таралуын бұзбауы тиіс.

Модель сапасын бағалау

Нақты жағдай / Болжам	**Шабуыл деп болжанды**	**Қалыпты деп болжанды**
Нақты шабуыл	TP — дұрыс анықталған шабуыл	FN — анықталмай қалған шабуыл
Нақты қалыпты	FP — жалған ескерту	TN — дұрыс анықталған қалыпты трафик

$$Precision = TP / (TP + FP).$$

Бұл көрсеткіш модель шабуыл деп белгілеген оқиғалардың қаншасы шынымен шабуыл екенін көрсетеді.

Recall = TP / (TP + FN) — нақты шабуылдардың қаншасы табылғанын көрсетеді.

F1-score = 2 × Precision × Recall / (Precision + Recall) — Precision мен Recall арасындағы теңгерімді сипаттайды.

Мысал. TP=80, FP=20 және FN=10 болсын. Precision=80/(80+20)=0,80. Recall=80/(80+10)=0,889. F1-score шамамен 0,842 болады. Бұл жағдайда модель нақты шабуылдардың көп бөлігін анықтайды, бірақ шабуыл деп берілген ескертулердің шамамен 20%-ы жалған.

Шекті мәнді таңдау

Көптеген модельдер бірден «қалыпты» немесе «шабуыл» класын емес, ықтималдық немесе аномалия ұпайын береді. Мысалы, 0 мен 1 арасындағы ұпай 0,82 болуы мүмкін. Егер шекті мән 0,5 болса, объект шабуыл деп белгіленеді. Шекті мәнді төмендету Recall көрсеткішін жоғарылатуы мүмкін, бірақ FP саны да артады. Керісінше, жоғары шек жалған ескертуді азайтады, алайда кейбір шабуылдар анықталмай қалуы мүмкін. Сондықтан шек техникалық метрикалармен ғана емес, ұйымның тәуекеліне қарай таңдалады.

Практикалық мысал

Ұйымдағы бір жұмыс станциясы қалыпты жағдайда сағатына 20–30 сыртқы желілік байланыс орнатады делік. Бір күні осы құрылғы бес минут ішінде жүздеген әртүрлі сыртқы мекенжайға қысқа байланыс жасай бастады. Connection rate, unique destination count және session duration белгілері қалыпты профильден қатты ауытқиды. Аномалия детекторы оқиғаға жоғары ұпай береді.

Бұл жерде жүйе құрылғыны бірден «шабуылшы» деп жарияламауы тиіс. SOC маманы қосымша контекстті қарайды: осы уақытта бағдарламалық жаңарту болды ма, рұқсат етілген қауіпсіздік сканері жұмыс істеді ме, құрылғыда жаңа процесс пайда болды ма, DNS және endpoint журналдарында қандай оқиғалар бар. Егер заңды себеп табылмаса және бірнеше дереккөз күмәнді белсенділікті растаса, оқиға инцидент ретінде жоғары басымдыққа ие болады.

Concept drift және модельді жаңарту

Ақпараттық жүйе тұрақты емес. Жаңа сервистер қосылады, пайдаланушылардың жұмыс тәртібі өзгереді, бұлттық ресурстар енгізіледі және трафик көлемі өседі. Соның салдарынан оқыту кезінде қалыптасқан «қалыпты» профиль бірнеше айдан кейін ескіруі мүмкін. Деректер таралуының осындай өзгерісі concept drift деп аталады.

Concept drift жалған ескертулер санын көбейтуі немесе жаңа қалыпты белсенділікті шабуыл ретінде көрсетуі мүмкін. Сондықтан модельдің Precision, Recall, F1-score және ескерту статистикасы уақыт бойынша бақыланады. Қажет жағдайда модель жаңа тексерілген деректермен қайта оқытылады. Модель нұсқалары мен оқыту деректерінің мерзімі құжатталуы тиіс.

Машиналық оқытудың артықшылықтары мен шектеулері

Критерий	**Артықшылығы**	**Шектеуі**
Белгісіз қауіптер	Қалыпты профильден жаңа ауытқуларды табуы мүмкін	Аномалияның бәрі шабуыл емес
Үлкен деректер	Көп көлемдегі трафикті автоматты талдайды	Есептеу және сақтау ресурстары қажет
Бейімделу	Жаңа деректермен қайта оқытуға болады	Concept drift тұрақты бақылауды талап етеді
Автоматтандыру	SOC ескертулерін басымдықтауға көмектеседі	Толық автоматты шешім қате әрекетке әкелуі мүмкін
Күрделі заңдылықтар	Адамға байқалмайтын байланыстарды анықтай алады	Кейбір модельдердің шешімін түсіндіру қиын

Қорытынды

Машиналық оқыту желілік аномалияларды анықтауда қалыпты мінез-құлықты модельдеуге, үлкен көлемдегі трафикті автоматты талдауға және белгісіз ауытқуларды бөліп көрсетуге мүмкіндік береді. Тиімді жүйе тек алгоритм таңдаумен шектелмейді. Деректерді дұрыс жинау, белгілерді қалыптастыру, train/test бөлудің дұрыстығы, класс теңгерімсіздігін ескеру, Precision және Recall арасындағы теңгерім, шекті мәнді таңдау және concept drift мониторингі бірдей маңызды.

Интеллектуалды детектордың нәтижесі басқа қауіпсіздік көздерімен салыстырылып, контекстпен тексерілгенде ғана сенімді шешімге айналады. Сондықтан ML дәстүрлі IDS/IPS, журналдарды талдау және қауіпсіздік маманының сараптамасын толықтыратын құрал ретінде қарастырылады.

Бақылау сұрақтары

1. Желілік аномалия дегеніміз не және ол кибершабуылдан несімен ерекшеленеді?
2. Қалыпты желілік мінез-құлық профилі қалай қалыптастырылады?
3. Желілік трафикті сипаттайтын негізгі белгілерді атаңыз.
4. Бақыланатын және бақыланбайтын оқытудың айырмашылығы қандай?
5. Isolation Forest және LSTM қандай міндеттер үшін қолданылуы мүмкін?
6. Data leakage модельді бағалауға қалай әсер етеді?
7. Класс теңгерімсіздігі кезінде Accuracy неге жеткіліксіз?
8. Precision, Recall және F1-score нені көрсетеді?
9. Шекті мәнді өзгерту False Positive және False Negative санына қалай әсер етеді?
10. Concept drift дегеніміз не және модельді не үшін қайта бағалау қажет?