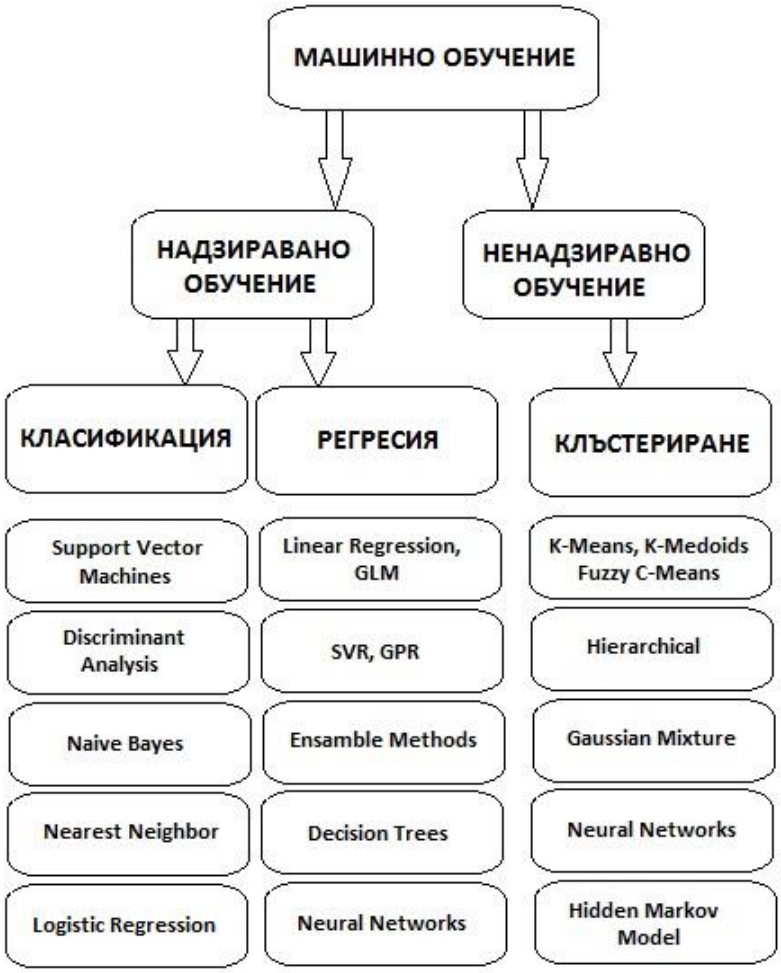
Видове алгоритми, използвани за машинно обучение

Основните подходи в машинното обучение са надзиравано (контролирано) машинно обучение (*Supervised*); ненадзиравано (неконтролирано) машинно обучение (*Unsupervised*). При контролираното машинно обучение има два типа задачи - регресионни задачи (*Regression problems*) и класификационни задачи (*Classification problems*). Неконтролирано машинно обучение решава задачи за клъстеризация (*Clustering problems*).

На Фигура 1 са дадени основните методи за машинно обучение и тяхното приложение. Проблемът за причисляването на даден инцидент към съответния тип, се отнася към задачите за класификация.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.



Фигура 1. Най-често ползвани методи за машинно обучение

Базова текстова класификация чрез методи на изкуствен интелект и машинно обучение е познат проблем в практиката. При анализа на научната литература, където се предлагат методи за решаване на проблема с класификацията на инцидентите в сигурността, се откроява основно използването на алгоритми за класификация чрез контролирано

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

обучение като „Метод на опорните вектори“ (*Support Vector Machine - SVM*) и „K-най-близък съсед“ (*K-Nearest Neighbor - KNN*). Освен тези се ползват бейсов класификатор и други модели базирани на дървета на решенията (*Decision Trees*), както и ансамблови алгоритми.

1. TF-IDF (Term Frequency-Inverse Document Frequency)

Много приложения използват TF-IDF като предварителен текст и подготовка на входа за алгоритми за класификация. Въз основа на статистически данни TF-IDF, определя честотата на поява на думи в текста на описанието на инцидента. Това помага да се извлече значението от текст, без да се налага да се прави пълен семантичен анализ на описанието. TF е броят пъти (честотата), в които даден термин се появява в документ. Теглото на даден термин, който се среща в документ, е пропорционално на честотата на термина. IDF (обратната честота на документа) намалява тежестта на термините, които се срещат много често в документа, и увеличава тежестта на термините, които се срещат рядко. Например думата „*the*“ е много често срещана, но не е добра ключова дума за включване в процес на класификация на текст.

2. Метод на опорните вектори (Support Vector Machine - SVM)

Причината, поради която SVM са подходящи за решаването на поставения проблем е, че този контролиран метод за машинно обучение директно генерира класификация и защото този метод е известен като добър за задачи за категоризиране на текст. С други думи, при SVM не е необходимо да се оценяват проби и да се конструира класификатор за тях, тъй като това е вградено в SVM процеса. Често този метод се ползва и при класифициране на софтуер като вредителски или не, но намира приложение и при класифицирането на доклади за инциденти по класификационна схема. Обикновено този метод се използва „за получаване

на класа за дадено описание на инцидента“ въз основа на текстова категоризация, която е от съществено значение за извличане на ключовите думи от доклад за инцидент.

Статистическите методи като TF-IDF и булево претегляне за представяне на текст създават по-малки входни вектори за SVM класификатор. Описанието на инцидента има важна роля при категоризирането на инцидентите. Както е показано при обучението, данните само с описанието на инцидента имат резултат от точност от 86% . Но обучението на данните само с номинални атрибути общата точност е 43%. Определено класификацията с атрибут за текстово описание дава много по-добри резултати.

1.3.3. K-най-близък съсед (K-Nearest Neighbor - KNN)

Това е метричен алгоритъм, при който се измерват разстояния между обектите за автоматична класификация. Алгоритъмът KNN използва евклидовото разстояние като показател за класифициране, за да се идентифицират кои са най-близките. Подходящ за решаването на класификационната задача, като може да се комбинира с TF-IDF при предварителната обработка за извличане на статистическа информация от текст на унифицирания доклад за инцидента.

На еднакъв набор от данни резултатите от KNN са малко по-лоши от резултатите по метода на SVM. Точността на атрибута на текстовото описание е 80%. Номиналните атрибути имат 42% точност. Ето защо една от най-важните части при решаването на класификационните проблеми е намирането на подходящи „характеристики“, такива които адекватно отразяват обективните зависимости от класификационния модел.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

1.3.4. Дърво на решенията (Decision Trees)

Друг интуитивен начин за класификация са дърветата с решения. Този метод е от тип контролирано машинно обучение. Дървото с решения може да се използва както за регресия, така и за проблеми с класификация. „Основната концепция на този модел е разделяне на множеството от данни на по-малки множества от данни, които са концептуализирани върху описателните характеристики. Това разделяне се извършва, докато се получи най-малкото множество от данни, съдържащо елементите от данни, които да бъдат категоризирани под един и същ етикет” . Това са двоични дървета, които се състоят от възли и листа. На всеки възел се взема решение въз основа на стойността на една от характеристиките, по които се класифицира. Ако функцията е непрекъсната, решението се основава на въпроса дали дадена стойност е по-голяма или по-малка от зададената. Преминавайки през дървото надолу и достигайки до листо, класификацията е завършена, понеже всяко листо е свързано с клас. Прилагането на този подход може да помогне при класифицирането на получените етикети към класификационна схема - различните видове инциденти, са отделни листа от дървото на решенията.

Налични са резултати, при които методът на дърветата за решения, комбиниран с TF-IDF, има около 90% точност. На същия набор от данни SVM произвежда подобна точност, но *Naive Bayes* представя различни резултати с 85% за TF-IDF и дори по-ниски с 55% само с терминална честота (TF).

1.3.5. Бейсова класификация

Naive Bayes заедно със SVM са добри алгоритми за класификация на текста. Класификаторът NB разчита на вероятностите от събития и се основава на теоремата на Байес . Вероятността за избор на категория на инцидента се изчислява за всички категории.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Този метод е добър за класификация на текст и документи. Класификаторът NB отчита броя на срещане на всяка дума и игнорира подреждането. Всеки „документ може да бъде представен като p-вектор на сумите (хистограма на честотата на думите)“. Той се използва за намиране на характеристики в докладите за инциденти, които принадлежат към определена категория. В литературата се използва т.н. бейсов метод за класификация на инциденти. Техният подход за класификация на инциденти е довел до 70% точност с 1000 характеристики.

Този тип класификация е често ползван при създаването на автоматизирани SPAM филтри, като постига много добри резултати.

1.3.6. Невронни мрежи

Използването на *Neural Network* за задачи по класификация е друга опция. Използва се *Softmax* класификатор, базиран на регресионна невронна мрежа. Проблемът с регресията е обобщен за класификационни проблеми, при които етикетът на класа може да приеме повече от две възможни стойности. Те сравняват точността, като използват само заглавието на билета или заглавието и описанието на билета. Отново резултатите са по-добри, ако се използва полето с текстово описание ~ 83%, спрямо ~ 80% само за заглавието. След това те сравняват NB и NN. Когато двете категории са избрани едновременно (заглавие и описание на билета), по-добрата обща точност от 85,8% постига NB класификатор.

В таблица 2 са обобщени най-често използваните алгоритми за машинно обучение за класификация на инциденти в киберсигурността въз основа на проучени статии и източници.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Класификационен метод	Вид метод – приложение	Изводи
Support Vector Machine SVM	Supervised ML – Classification	Подходящ за класификация на текстови документи
K-Nearest Neighbor - KNN	Supervised ML – Classification	Лесен за имплементиране и добър за класификация на текстови документи
Naive Bayes	Supervised ML – Classification	Добър за класификационни задачи и текстови документи
Decision Trees	Supervised ML – Regression/Classification	Добър за регресионни и класификационни задачи
Neural networks	Supervised ML – Regression→Classification	Подходящ за намиране на аномалии и класификационни проблеми

Таблица 1. Алгоритми за машинно обучение за класификация на инциденти в киберсигурността

Всички те са контролирани методи за машинно обучение, подходящи за решаване на класификационни проблеми.

Невронните мрежи и дълбокото обучние обикновено дават много добри резултати за огромни масиви от данни, от порядъка на милиони записи. Базата данни CVE (Common Vulnerabilities and Exposures) с описание на инциденти в киберсигурността, която ще използваме, е сравнително малка по обем, с около сто хиляди записа, което прави класическите методи за машинно обучение по-подходящи за решаването на класификационната задача.

Унифицираните доклади за инциденти обикновено се отнасят до конкретна организация и имат по-частен характер. От друга страна, базата данни Common

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Vulnerabilities and Exposures (CVE) съдържа информация за всички новооткрити уязвимости в глобален мащаб. По този начин, адекватното използване на методи за изкуствен интелект при класификацията на елементите от CVE базата данни според типа на уязвимост би била обобщен случай на класификацията на доклади за инциденти възникнали в рамките на дадена организация.

Класификацията и прогнозирането на уязвимостите са актуална област на изследване и привлича все по-голямо внимание. Някои автори като са фокусирани върху отстраняването на уязвимостите и прогнозирането на експлоатацията на уязвимостите. Други предлагат нови модели за прогнозиране на експлоатацията, които подобряват базовия модел. Авторите се фокусират върху референтната информация (като URL адреси) в CVE за определяне на вероятността за експлоатиране на уязвимости. Други са изследвали, че „добавяйки общи думи и n-грами като характеристики, е възможно да се открият модели, които не са уловени от простите параметри на CVSS и CWE от NVD“ . Това прави класификацията по-добра. Използва се техника за обединяване на характеристики по време на тренирането на моделите си, за да се направи по-добро прогнозиране за експлоатация на разкритите уязвимости в софтуера. Приоритизирането на софтуерните уязвимости е друга основна тема в CVE класификацията. Ръчното оценяване на уязвимостите може да причини проблеми като забавяне или задаване на ниски CVSS оценки за по-тежки уязвимости. В (Wright, [131]2019) се вижда, че прогнозирането с машинното обучение използвано за оценка на уязвимостите, е по-бързо от традиционните подходи. Показано е, че при автоматизирания начин се постига по-висока точност, отколкото при нормалната CVSS система за оценка. Не всички CVE имат информация за CPE, продукт и доставчик. Това затруднява машинната обработка. Автори предлагат автоматичен процес на съпоставяне на описанието на CVE с CPE. Дълбокото обучение и невронните мрежи често се използват за откриване на софтуерни уязвимости в кода, „описващ разликата между уязвима част от кода и неговата коригирана

версия“ подчертават идеята за използване на дълбоки невронни мрежи за подпомагане на анализа на двоичен код вместо изходен код. Авторите (Montuno, [134]2019) описват алгоритми за машинно обучение при оценка на уязвимостта, а (Zhou, Liu, Siow, Du, & Liu, [135]2019) предлагат графови невронни мрежи за ефективно идентифициране на уязвимостта чрез обучение върху семантика на софтуерната програмата.

В научната литература се откроява наблюдавана точност в интервала 75-90%, при прилагането на алгоритмите за машинна класификация като SVM, KNN и дърво на решенията. Това би било добре да се подобри, тоест да се намалят фалшиво положителни и фалшиво отрицателни резултати. Посоката на работа ще бъде въз основа на избран по определи критерии най-ефективен метод за конкретната задача, който се разглежда като базов, да бъдат приложени допълнителни методи за съответното му подобрение. Очаква се да се получи хибриден модел с по-добри характеристики.

Наличието на не голямо количество източници по темата може да се обясни и с това, че основно изследванията до сега са били съсредоточени към „тактическото киберразузнаване“, тоест при непосредственото отразяване на атаки, което вече не е предмет на научно-изследователска дейност, а е индустриална дейност, при която много системи са създадени и интегрирани и са в ежедневна употреба в практиката.

Но при другите два етапа от киберзащитата - „оперативното киберразузнаване“ и обработване на инциденти, задачите от вида „подпомагане на вземане на решения“, чрез методите на изкуствен интелект са в процес на изследване.

Като основен извод от проучената литература, става явно, че най-важната част от решаването на тази задача за класификация е намирането на така наречените „*features*“, т.е. характеристики, които отразяват адекватно обективни зависимости от класификационния статус. Извличането на тези класификационни характеристики („*features*“) е решаваща стъпка.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.