КОМПЬЮТЕРНОЕ ОБЕСПЕЧЕНИЕ И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

DOI: 10.24143/2072-9502-2020-1-29-40
УДК 004.7:004.056.5

АНАЛИЗ ПРИМЕНЕНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ КОМПЬЮТЕРНЫХ СИСТЕМ ДЛЯ ПОВЫШЕНИЯ ЗАЩИЩЕННОСТИ ОТ МОШЕННИЧЕСКИХ ТЕКСТОВ

С. Д. Шибайкин, В. В. Никулин, А. А. Аббакумов

Национальный исследовательский Мордовский государственный университет им. Н. П. Огарева,
Саранск, Республика Мордовия, Российская Федерация

Неотъемлемым условием функционирования каждой компании, работа которой связана с хранением информации, является безопасность в сфере IT. Проанализированы различные модели детекции мошеннических текстов, включая машину опорных векторов, нейронные сети, логистическую регрессию, наивный байесовский классификатор. Предложено повысить эффективность детекции мошеннических сообщений путем объединения классификаторов в ансамбли. Метаклассификатор позволяет учитывать значения точности всех анализаторов, задействуя в работе построение матрицы весов и характеристику, определяющую минимальную границу точности. На базе разработанного метода создан и апробирован программный модуль классификации мошеннических текстовых сообщений, написанный на языке Java с использованием класса M1 открытой библиотеки OPENCV. Приведен общий алгоритм работы ансамблевого метода. Выполненный эксперимент на базе логистической регрессии, наивного байесовского классификатора, многослойного персептрона и ансамбля этих классификаторов выявил максимальную эффективность наивного байесовского алгоритма классификации и перспективность объединения классификаторов в ансамбли. Комбинированные методы (ансамбли) улучшают результаты и увеличивают эффективность анализа в отличие от работы отдельных анализаторов.

Ключевые слова: мошеннический текст, детекция, текстовые данные, машинное обучение, классификатор, нейронная сеть, ансамбль-система, алгоритм.

Для цитирования*: Шибайкин С. Д., Никулин В. В., Аббакумов А. А.* Анализ применения методов машинного обучения компьютерных систем для повышения защищенности от мошеннических текстов // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2020. № 1. С. 29–40. DOI: 10.24143/2072-9502-2020-1-29-40.

Введение

В настоящее время исследование программных средств для борьбы с киберпреступлениями и мошенничеством приобретает особую актуальность. Безопасность в сфере IT считается неотъемлемым условием функционирования каждой фирмы, работа которой связана с хранением информации. На данный момент в подавляющем числе IT-компаний Internet используется как главная составляющая деятельности фирмы, что приводит к увеличению числа программных и аппаратных средств защиты от небезопасного функционирования систем [1–3]. И, несмотря на то, что сложно оценить экономическую выгоду от внедрения данных систем, актуальность этих мероприятий трудно переоценить.

Под мошенническим текстом понимается сообщение, составленное с целью введения собеседника в заблуждение. Задача детекции мошеннических текстов заключается в определении,

был ли текст написан человеком либо был создан при помощи программы. Для решения данных задач нужно учесть следующие особенности предметной области: в настоящее время наиболее актуальным считается использование машинного обучения, которое может выявить определенные закономерности в текстовых массивах данных. Текстовый майнинг (англ. text mining – получение информации из текста) – это широкая область исследований, которая завоевала популярность вместе с ростом объемов текстовых данных. Автоматизация ряда прикладных задач, таких как анализ содержания текста, классификация и каталогизация документов, обобщение текста, машинный перевод, в настоящее время часто выполняется с использованием моделей машинного обучения. Детекция мошеннических текстов представляет собой классический пример задачи классификации документов, который включает в себя классификацию потоков текстовых данных (например, электронной почты) как нежелательного контента.

Для постановки эксперимента мы выполним следующую последовательность шагов:

– подготовка текстовых данных;

– создание словаря;

– процесс извлечения характерных черт (Feature extraction);

– выбор и обучение классификаторов;

– оценка эффективности результатов классификации;

– вывод об эффективности процесса детекции мошеннических текстов на основе проведенной классификации.

Для того чтобы провести оценку точности и эффективности анализа больших объемов данных средствами машинного обучения, следует использовать методики, которые основаны на вычислении матриц ошибок и расчетах, проведенных на основе метрик. Матрицы ошибок (confusion matrix) применяют в тех случаях, когда необходимо представить формализованное описание качества применяемых аналитических моделей. Практически всегда для них уже рассчитаны численные параметры соответствий классов и объектов.

Вычисление показателей эффективности ансамбль-системы

Ансамбль-система – это комбинация различных методов машинного обучения, объединенных в один классификатор. Задача ансамбль-систем состоит в объединении набора анализаторов в одну общую систему, вследствие этого достигается повышение точности анализа данных. Главными элементами ансамбль-систем являются моноклассификаторы и метаклассификаторы. Для оценки эффективности детекции мошеннических текстов с помощью ансамбль-систем будет выполнен анализ тестовых данных. Выполним вычисление показателей эффективности ансамбль-системы, полученные значения объединим в вектор эффективности системы. Данный шаг позволит выполнить анализ и сопоставление результатов эффективности детекции классификаторов и системы.

Проведенный расчет дает возможность вычислить следующие значения характеристик отдельных нейросетей и ансамбль-системы:

– первая нейросеть, отличающаяся максимальной мощностью нейронного слоя, будет использовать первый тестовый набор данных;

– вторая нейросеть, имеющая среднюю мощность нейронного слоя, будет также использовать первый тестовый набор данных;

– третья нейросеть, которая отличается пониженной мощностью нейронного слоя, будет, как первая и вторая, использовать первый тестовый набор данных;

– четвертая нейросеть, как и вторая, имеет среднюю мощность нейронного слоя, будет использовать второй тестовый набор данных;

– ансамбль-система соединяет четыре вышеописанные нейронные сети с помощью обученного классификатора.

По результатам анализа значений метрик эффективности включенных в систему классификаторов, а также самой ансамбль-системы были сделаны следующие заключения:

– точность решения, которое было спрогнозировано ансамбль-системой, схожа с показателем точности самого эффективного анализатора. В некоторых случаях данная точность ансамбль-системы превышает точность классификаторов. Данная характеристика основана на главных принципах построения классификаторов. Данный факт был подтвержден экспериментально;

– ошибочный прогноз, сделанный ансамбль-системой, равен показателю ошибочности классификаторов, однако в некоторых случаях он может быть выше. Несмотря на это, применение ансамбль-системы в любых случаях помогает не допустить серьезных ошибок в классификации, если анализаторы дают неверные прогнозы относительно принадлежности объектов к заданным классам;

– в случае если все анализаторы системы имеют наивысший показатель ошибочности, т. е. неверно относят объекты к нужным классам, сама ансамбль-система будет наследовать эту особенность. Экспериментально доказано, что данная неточность не может быть выше неточности самого неэффективного анализатора в системе. В теории эта особенность объясняется присутствием верного подхода системы к выбору финальной гипотезы.

Описанные выше выводы также экспериментально подтверждаются при работе с другими тестовыми данными. Актуальность применения ансамбль-систем для детекции мошеннических текстов обуславливается большими объемами тестовых данных, множеством вариаций результатов, сложностью подбора эффективных инструментов для решения данной задачи. Использование ансамбль-систем эффективно решает данные вопросы, а также снижает показатели ошибочности и повышает точность детекции мошеннических текстов. Для эффективной реализации системы детекции мошеннических текстов будут применены различные методики: машина опорных векторов, искусственные нейросети, деревья решений. Однако в настоящий момент нет эффективного инструмента, решающего поставленную задачу. Методы, которые комбинируют несколько способов (ансамбли), улучшают результаты и увеличивают эффективность анализа в отличие от работы отдельных анализаторов. Бутстрэп-агрегирование, или бэггинг, – это алгоритм, комбинирующий результаты классификации нескольких модулей.

Бустинг (англ. boosting – усиление), в отличие от бэггинга, работает следующим образом. Для каждого вектора системы определяется вес. Классификаторы выполняют несколько итераций: после определения веса тестовых данных на каждом следующем построении классификаторов отделяются данные, распознанные неверно. Конечное решение выносится голосованием, коэффициенты весов являются функцией точности анализаторов. В настоящий момент имеется множество методик их подбора, одним из самых эффективных остается алгоритм Adaboost.

Стекинг – алгоритм агрегации различных анализаторов. В отличие от двух предыдущих алгоритмов стекинг применяется для соединения модулей, работающих по разным алгоритмам. Тестовый набор обучающих данных делится на следующие категории: в первой анализаторы, работающие в системе, проходят обучение на первом наборе данных, а тестируются на втором, затем, после проведения анализа результатов, будет спроектирован метаклассификатор, который принимает решение.

Описанная выше ансамбль-система по аналогии с алгоритмом стекинга будет агрегировать различные анализаторы. Однако в отличие от существующих описанных выше алгоритмов агрегации метаклассификатор будет учитывать значения точности всех анализаторов, задействуя в работе построение матрицы весов. Также будет задействована характеристика, определяющая минимальную границу точности, с помощью которой проводится классификация, т. е. определяется принадлежность объектов к классам.

Сравнительный анализ алгоритмов

Приведем сравнительный анализ описанных выше алгоритмов с целью выявления лучшего решения для повышения точности детекции мошеннических текстов. Алгоритмы с различными подходами можно использовать в роли анализаторов для построения ансамбль-системы. Для проведения опыта были выбраны модели, которые смогут эффективно детектировать мошеннические тексты: машина опорных векторов, нейросети и деревья решений. Для проведения эксперимента (рис. 1) нужно выполнить следующие шаги:

– определить тестовые данные для классификации;

– обучить и протестировать классификаторы на тестовом наборе данных.

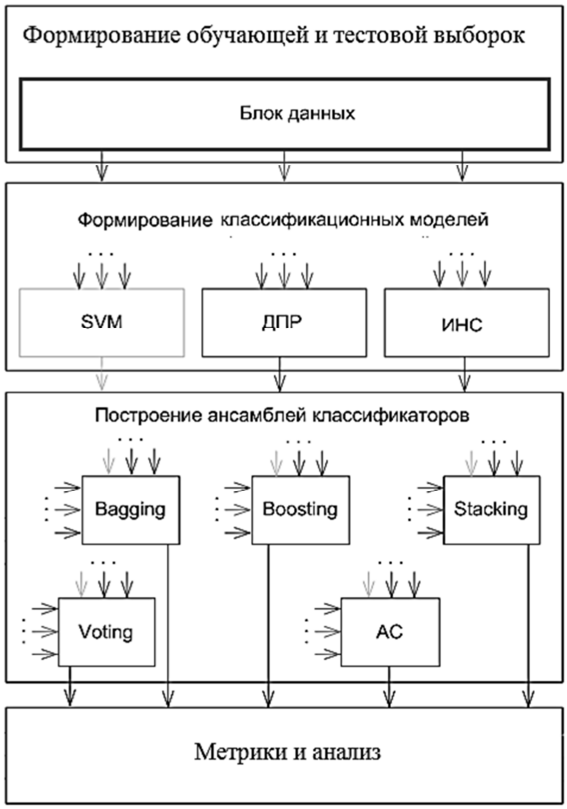

Рис. 1. План проведения эксперимента по оценке точности ансамбль-систем:
SVM – алгоритм машинного обучения Support Vector Machines; ДПР – дерево принятия решений;
ИНС – искусственная нейронная сеть

Support Vector Machines (SVM) является одним из лучших контролируемых алгоритмов машинного обучения для построения бинарного классификатора. Данный метод сводит обучение классификатора к решению эвристическими алгоритмами путем последовательного уменьшения целевой функции (рис. 2).

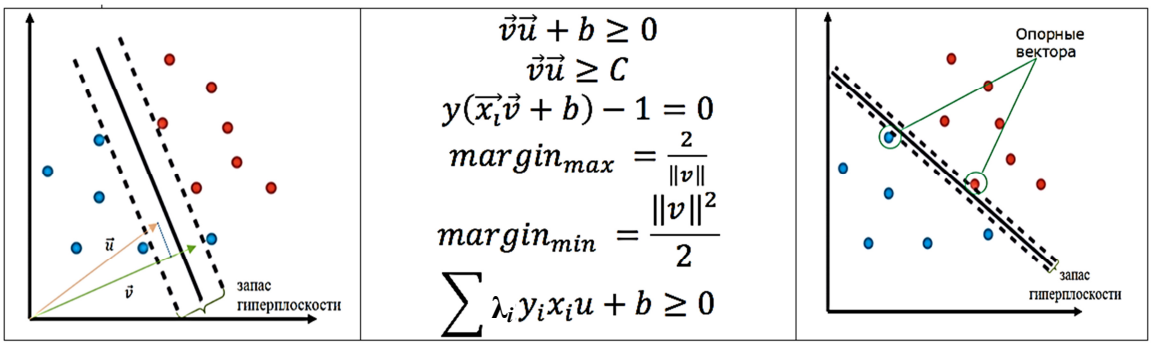

Рис. 2. Параметры SVM: с широкой разделяющей плоскостью (*а*);
с узкой разделяющей плоскостью (*б*)

Сотрудниками кафедры инфокоммуникационных технологий и систем связи (ИКТСС) был проведен эксперимент, в ходе которого сформировали и обучили машину опорных векторов с различными ядрами: линейным (LINEAR), RBF, сигмоидальным (SIGMOID), полиномиальным (POLY). Затем была проведена классификация текстов на предмет того, являются ли они мошенническими.

Параметр гамма (используется в ядрах POLY/RBF/SIGMOID) контролирует, насколько чувствительна граница к выбросам. Данный параметр определяет, насколько далеко распространяется влияние обучающей выборки. Если параметр гамма имеет низкое значение, то это означает, что каждый образец обучающей выборки имеет большой радиус влияния (имеет вес, достаточный для влияния). И, наоборот, при высоких значениях гамма каждый образец обучающей выборки имеет меньший радиус влияния, тем самым уменьшая вес точек, отдаленных от пороговой гиперплоскости.

Параметр *C* (используется в задачах) имеет отношение к геометрии границы, он обеспечивает компромисс между гладкой границей решения (сложностью модели) и точностью классификации тестовой выборки. Фактически параметр *C* обратно пропорционален дисперсии: чем он выше, тем меньше будет запас у гиперплоскости при условии, что эта гиперплоскость лучше справится с правильной классификацией обучающей выборки. Меньшее значение *C* оставляет больший запас, даже если эта плоскость имеет меньшую точность классификации. Данный алгоритм неустойчив к выбросам, поэтому для эффективной работы алгоритма необходимо управлять параметрами *C* и гамма, т. е. необходимо найти баланс между гиперплоскостью с наибольшим минимальным запасом и гиперплоскостью, которая более точно классифицирует обучающую выборку. В случае линейно разделимой выборки алгоритм сводится к максимизации ширины запаса.

Результирующие данные, сведенные в табл. 1, показывают, что наиболее эффективной стала машина опорных векторов, основанная на линейном ядре. Интересно, что машины этого типа более просты и быстрее обучаются.

Таблица 1

Метрики качества детекции мошеннических текстов с помощью машины опорных векторов с различными ядрами

Метрика точности	Support Vector Machine			
	Poly	Sigmoid	RBF	Linear
accuracy	0,958008117	0,963315642	0,971198876	0,988058071
TN	12238	12151	12222	12199
FP	3	90	19	42
FN	535	380	350	111
TP	36	191	221	460
sensitivity	0,063047285	0,334500876	0,38704028	0,805604203
specifity	0,999754922	0,99264766	0,998447839	0,996568908
Precision	0,923076923	0,679715302	0,920833333	0,916334661
npv	0,958114773	0,969675205	0,972160356	0,990982941
inf	0,062802207	0,327148535	0,385488119	0,802173111
mark	0,881191696	0,649390508	0,89299369	0,907317602
fl	0,118032787	0,448356808	0,545006165	0,857409133
miss	0,936952715	0,665499124	0,61295972	0,194395797
fallout	0,000245078	0,00735234	0,001552161	0,003431092
falseDiscovery	0,076923077	0,320284698	0,079166667	0,083665339
falseOmission	0,041885227	0,030324795	0,027839644	0,009017059

На рис. 3 также отчетливо видно превосходство SVM машины с линейным ядром.

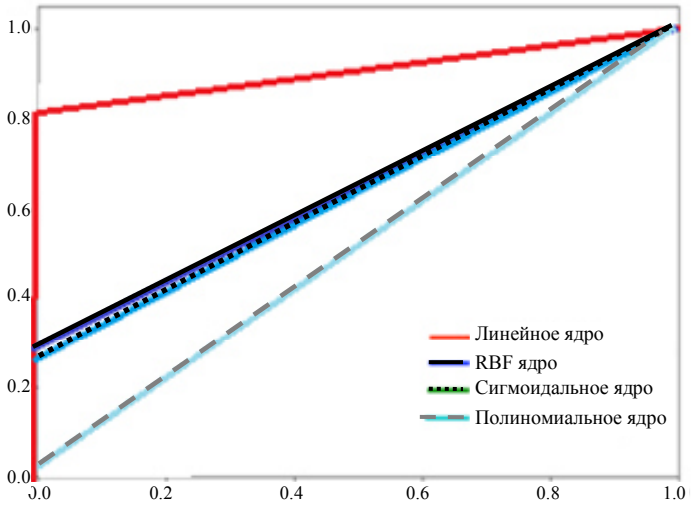

Рис. 3. ROC-кривые для машин опорных векторов с разными ядрами

Логистическая регрессия (Logistic Regression) является вероятностным методом многоклассовой классификации. Метод использует логистическую функцию (сигмоида – для описания вероятности того, что выборка принадлежит одному из двух классов) для преобразования входного значения критерия признака в прогнозируемое значение. Задача логистической регрессии оптимизировать параметр θ для достижения гипотезы правдоподобия $0 \le h_\theta \le 1$:

$$h_\theta(X) = \frac{1}{1+e^{-\theta^T X}},$$

где h_θ – гипотеза правдоподобия; X – категориальный признак.

В случае бинарной классификации, если $h_\theta < 0{,}5$, класс определяется как 0, иначе как 1. Подбор корректных параметров в алгоритме позволит уменьшить ошибки обучения и обеспечить высокую точность.

Параметр «скорость обучения» алгоритма определяет, насколько быстро необходимо обучить алгоритм. Чем выше значение параметра скорости обучения, тем выше скорость обучения, однако при этом снижается точность. Процесс обучения модели сводится к выбору формы сигмоиды, форма которой наилучшим образом соответствует нашим данным. Для предотвращения переобучения модели выполняется $L1$-регуляризация или $L2$-регуляризация. $L1$-регуляризация способствует разреженности функции, когда лишь немногие факторы равны нулю, путем отбора наиболее важных факторов, влияющих на результат. Он добавляет к функционалу потерь сумму модулей весов линейной модели. $L2$-регуляризация способствует появлению малых весовых коэффициентов модели путем запрета на непропорционально большие коэффициенты, но не способствует их точному равенству 0. Он добавляет к функционалу потерь сумму квадратов весов линейной модели с множителем λ.

Наивный байесовский классификатор (НБК) является наиболее простым вероятностным классификатором, использующим в своей основе теорему Байеса. Формула

$$P(C|F_1, \ldots, F_n) = \frac{P(C)P(F_1, \ldots, F_n|C)}{P(F_1, \ldots, F_n)}$$

позволяет рассчитать апостериорную вероятность $P(C|F_1, \ldots, F_n)$ на основе $P(C)$ – априорной (безусловной) вероятности класса C, $P(F_1, \ldots, F_n|C)$ – вероятности данных значений признаков при данном классе, $P(F_1, \ldots, F_n)$ – априорной вероятности данных значений признаков.

Классификатор предполагает, что наличие (или отсутствие) определенного признака класса не связано с наличием (отсутствием) какого-либо другого признака (модель независимых признаков).

Для построения НБК требуется вычисление математического ожидания и дисперсии по каждому признаку. Предположение о независимости признаков позволяет перейти от оценки *n*-мерной плотности к оценке *n* одномерных плотностей. Все параметры модели (априорные вероятности классов и распределение вероятностей признаков) могут быть аппроксимированы их относительными частотами в обучающей выборке. В случае обучения НБК на вещественных значениях (некатегорийные признаки) для классификации необходимо использовать гауссовское нормальное распределение

$$P\left(x = m|C\right) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(m-\mu_C)^2}{2\sigma_c^2}},$$

где μ_C – математическое ожидание класса C; σ_c^2 – дисперсия класса C; m – критерий гипотезы.

Наивный байесовский классификатор объединяет модель независимых признаков с решающим правилом. Одним из общих решающих правил является выбор гипотезы, которая является наиболее вероятной, а классификатор определяется функцией

$$Classificator\left(f_1, \ \ldots, f_n\right) = \operatorname{argmax} p\left(C = c\right) \prod_{i=1}^{n} P(F_i = f_i \mid C = c).$$

Бустинг – класс методов машинного обучения, основанный на идее комбинации простых классификаторов, полученных с помощью алгоритма обучения, способного классифицировать лучше, чем случайное угадывание. Основная идея бустинга состоит в последовательном обучении слабых классификаторов, каждый из которых пытается исправить ошибку предыдущих предикторов. Основными алгоритмами бустинга являются AdaBoost и Gradient Boosting. Общий принцип работы AdaBoost схож с Random Forest, т. к. оба объединяют прогнозы, сделанные каждым деревом решений для принятия решения об окончательной классификации. Отличием двух ансамблевых методов являются глубина дерева решений и влияние прогноза, сделанного каждым деревом, по-разному влияющие на окончательный прогноз.

Общий алгоритм состоит из следующих этапов:

– инициализация весов обучающей выборки. На данном этапе можно указать значимость элементов обучающей выборки. Изначально все элементы имеют одинаковый вес и в сумме на каждом шаге выбора слабого классификатора должны быть равны 1;

– выбираем слабый классификатор h_i. Для каждого признака создаем дерево решений с глубиной 1, после чего сравниваем полученные прогнозы с фактическими метками обучающей выборки. Наилучшее деление, которое выполнило классификацию, будет являться следующим деревом в лесу. Общая ошибка e_i слабого классификатора h_i складывается из суммы весов некорректно классифицированных образцов. Фактически наилучшее деление дерева решений обеспечивает минимизацию взвешенной ошибки e_i слабого классификатора h_i. Далее вычисляем вес α_i полученного слабого классификатора h_i. Чем больше значение α_i, тем большее влияние слабый классификатор h_i оказывает на целевую функцию $H(x)$;

– обновляем значение весовых коэффициентов обучающей выборки с учетом некорректно классифицированных образцов. Z_i – нормирующий параметр, равный сумме всех весовых коэффициентов i-го слабого классификатора;

– если $e_i \geq 0{,}5_i$, то останавливаем процесс, в противном случае повторяем процесс выбора слабого классификатора на обучающей выборке с новыми весовыми коэффициентами. Итоговая целевая функция имеет вид $H\left(x\right) = \operatorname{sign}\left(\sum_{i=1}^{M} \alpha_i h_i\left(x\right)\right)$. В случае отсутствия веса α_i целевая функция сводится к простому голосованию, как Random Forest.

Многослойный персептрон (MPL) является наиболее часто используемым типом искусственных нейронных сетей. Многослойный персептрон состоит из входного слоя, одного или нескольких

скрытых слоев и выходного слоя. Каждый слой (кроме входного) многослойного персептрона содержит в себе один или несколько нейронов, направленно связанных с нейронами из предыдущего и следующего (за исключением выходного слоя) слоев. На рис. 4 представлен четырехслойный персептрон с двумя входами, двумя выходами и двумя скрытыми слоями с 9-ю нейронами.

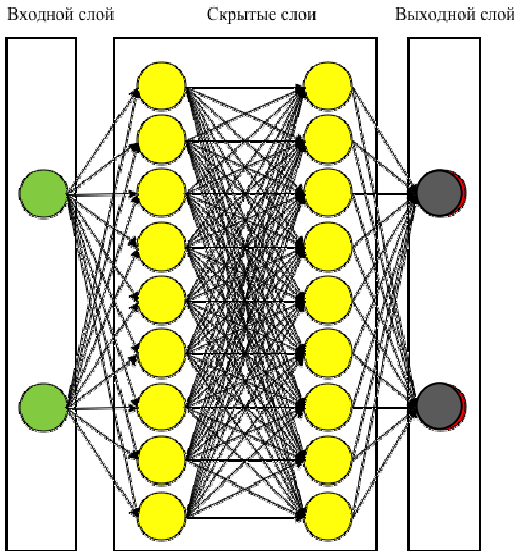

Рис. 4. Четырехслойный персептрон

Все нейроны в многослойном персептроне похожи между собой, они имеют несколько входных ссылок, которые принимают выходные значения от нейронов предыдущего слоя, и несколько выходных ссылок, которые передают ответы к нейронам следующего слоя. Значения, полученные из предыдущего слоя, суммируются с индивидуальными весами каждого нейрона и плюс значение смещения. Полученная сумма преобразуется с использованием активационной функции. Активационная функция (функция активации) определяет выходной нейрон.

Выполненный на кафедре ИКТСС эксперимент с тремя отмеченными выше классификаторами и ансамбль-системой выявил в итоге результаты, приведенные в табл. 2 и 3.

Таблица 2

Метрики качества детекции мошеннических текстов с помощью логистической регрессии, НБК, MPL, Ensembles

Метрика точности	Классификатор			
	Логистическая регрессия	Наивный байесовский классификатор	Многослойный персептрон	Ансамбль классификаторов
accuracy	0,988058071	0,992038714	0,990321574	0,991336247
TN	12199	12209	12232	12220
FP	42	31	8	20
FN	111	71	116	91
TP	460	501	456	481
sensitivity	0,805604203	0,875874126	0,797202797	0,840909091
specifity	0,996568908	0,99746732	0,999346405	0,998366013
precision	0,916334661	0,941729323	0,982758621	0,96007984
npv	0,990982941	0,994218241	0,990605766	0,992608237
inf	0,802173111	0,873341446	0,796549202	0,839275104
mark	0,907317602	0,935947564	0,973364387	0,952688077
f1	0,857409133	0,907608696	0,88030888	0,896551724
miss	0,194395797	0,124125874	0,202797203	0,159090909
fallout	0,003431092	0,00253268	0,000653595	0,001633987
falseDiscovery	0,083665339	0,058270677	0,017241379	0,03992016
falseOmission	0,009017059	0,005781759	0,009394234	0,007391763

Таблица 3

Метрики качества детекции мошеннических текстов с помощью различных методов обучения

Модель обучения	Классификатор	Точность	Время	
			Время обучения	Время анализа
Бэггинг	ИНС	55,38807	1,8931	0,0132
	SVM	49,93362	0,7865	0,0528
	ДПР	58,74345	0,0187	0,018
Бустинг	ИНС	56,85399	3,5343	0,0156
	SVM	56,01708	0,9185	0,0552
	ДПР	56,01906	0,0308	0,0216
Стэкинг	ИНС	50,09859	0,3069	0,0228
	SVM	50,09427	0,1804	0,0168
	ДПР	50,09886	0,0132	0,0204
Голосование	ИНС, SVM, ДПР	59,16006	0,5434	0,0648
Ансамбль-система	ИНС, SVM, ДПР	61,974	0,5896	0,0612

Максимальную эффективность выявил наивный байесовский классификатор. Он же оказался более результативным, чем машина опорных векторов.

Построенные ROC-кривые иллюстрируют максимальную эффективность наивного байесовского алгоритма классификации (рис. 5).

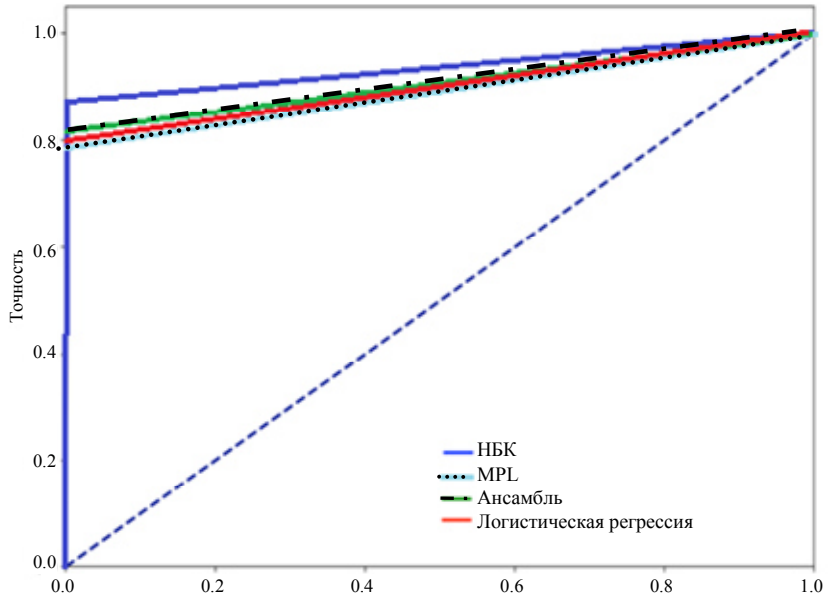

Рис. 5. ROC-кривые для логистической регрессии, наивного байесовского алгоритма классификации, многослойного персептрона и ансамбля этих классификаторов

Несмотря на высокую эффективность наивного байесовского алгоритма классификации эксперимент показал перспективность объединения классификаторов в ансамбли с целью повышения точности детекции мошеннических текстов (см. табл. 3).

С точки зрения временных затрат спроектированная ансамбль-система показывает средний результат [4], в ряде случаев опережая системы, созданные на базе бэггинга и бустинга. Во всех случаях ансамбль-система обучалась и классифицировала объекты в рамках конечного предсказуемого времени.

Полученные теоретические результаты позволяют более осознанно и эффективно подходить к решению задачи детекции мошеннических текстов, а написанный программный модуль на языке Java с использованием класса Ml открытой библиотеки OPENCV позволяет эффективно выполнять классификацию мошеннических текстов.

Заключение

В работе проанализирована проблемная ситуация детекции мошеннических текстов. Детерминированы виды мошеннических текстов, представлены техники маскировки мошеннических текстов. Проведена оценка эффективности результатов классификации. В дополнение к этому эксперимент показал перспективность объединения классификаторов в ансамбли с целью повышения точности детекции мошеннических текстов. При проектировании системы создаются и обучаются анализаторы, входящие и функционирующие в составе метаклассификатора, который отвечает за их комбинацию в одно целое, а также оценивает эффективность детекции. На основании вышеизложенного можно сделать вывод о том, что сравнительный анализ машинных методов обучения показал перспективность их использования не только для детекции мошеннических текстов, но и для классификации ИТ-сервисов и инцидентов библиотеки ITIL, которая позволит эффективно маршрутизировать пользовательские заявки.

СПИСОК ЛИТЕРАТУРЫ

1. *Красоткин М. А., Шибайкин С. Д.* Исследование программно-аппаратных средств для борьбы с мошенничеством // Прикладная математика и информатика: современные исследования в области естественных и технических наук: сб. науч. ст. IV Науч.-практ. междунар. конф. (школы-семинара) молодых ученых (Тольятти, 23–25 апреля 2018 г.): в 2 ч. Тольятти: Изд-во Качалин Александр Васильевич, 2018. Ч. 2. С. 162–168.

2. *Красоткин М. А., Шибайкин С. Д.* Алгоритмы борьбы с мошенничеством в телекоммуникационных системах // Материалы XXII Науч.-практ. конф. молодых ученых, аспирантов и студентов национ. исследоват. Мордов. гос. ун-та им. Н. П. Огарева (Саранск, 25 сентября–01 октября 2018 г.): сб. тр.: в 3 ч. Саранск: Изд-во Мордов. гос. ун-та им. Н. П. Огарева, 2019. Ч. 1. С. 148–151.

3. *Ладанова Е. О., Никулин В. В.* К вопросу об информационной безопасности при анализе текстовых сообщений // Вопросы информационной безопасности: материалы II Межрегион. вебинара (Саранск–Елец, 21 февраля 2018 г.). Саранск: Изд-во Мордов. гос. ун-та им. Н. П. Огарева, 2018. С. 59–63.

4. *Ямашкин С. А.* Методическое и алгоритмическое обеспечение процесса анализа структуры земель на базе данных дистанционного зондирования: дис. … канд. техн. наук. Саранск, 2016. 186 с.

Статья поступила в редакцию 18.09.2019

ИНФОРМАЦИЯ ОБ АВТОРАХ

Шибайкин Сергей Дмитриевич – Россия, 430005, Саранск; Национальный исследовательский Мордовский государственный университет им. Н. П. Огарева; канд. техн. наук; доцент кафедры инфокоммуникационных технологий и систем связи; shibaikinsd@mail.ru.

Никулин Владимир Валерьевич – Россия, 430005, Саранск; Национальный исследовательский Мордовский государственный университет им. Н. П. Огарева; канд. техн. наук, доцент; зав. кафедрой инфокоммуникационных технологий и систем связи; nikulinvv@mail.ru.

Аббакумов Андрей Александрович – Россия, 430005, Саранск; Национальный исследовательский Мордовский государственный университет им. Н. П. Огарева; канд. техн. наук; доцент кафедры автоматизированных систем обработки информации и управления; abbakumov_aa@mail.ru.

ANALYSIS OF MACHINE LEARNING METHODS FOR COMPUTER SYSTEMS TO ENSURE SAFETY FROM FRAUDULENT TEXTS

S. D. Shibaikin, V. V. Nikulin, A. A. Abbakumov

National Research Ogarev Mordovia State University,
Saransk, Republic of Mordovia, Russian Federation

Abstract. IT Security is an essential condition for functioning of each company whose work is related to the information storage. Various models for detecting fraudulent texts including a support vector machine, neural networks, logistic regression, and a naive Bayes classifier, have been analyzed. It is proposed to increase the efficiency of detection of fraudulent messages by combining classifiers in ensembles. The metaclassifier allows to consider the accuracy values of all analyzers, involving in the work the construction of the weight matrix and the characteristic that determines the minimum accuracy boundary. Based on the developed method, a software module for the classification of fraudulent text messages written in Java using M1 class of the OPENCV open library was created and tested. The general algorithm of the ensemble method is given. An experiment based on logistic regression, a naive Bayesian classifier, a multilayer perceptron, and an ensemble of these classifiers has revealed the maximum efficiency of the naive Bayesian classification algorithm and the prospect of combining classifiers into ensembles. The combined methods (ensembles) improve the results and increase the efficiency of the analysis, in contrast to the work of individual analyzers.

Key words: fraudulent text, detection, text data, machine learning, classifier, neural network, ensemble-system, algorithm.

For citation: Shibaikin S. D., Nikulin V. V., Abbakumov A. A. Analysis of machine learning methods for computer systems to ensure safety from fraudulent texts. *Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and Informatics*. 2020;1:29-40. (In Russ.) DOI: 10.24143/2072-9502-2020-1-29-40.

REFERENCES

1. Krasotkin M. A., Shibaikin S. D. Issledovanie programmno-apparatnykh sredstv dlia bor'by s moshennichestvom [Study of hardware and software for fraud fighting]. *Prikladnaia matematika i informatika: sovremennye issledovaniia v oblasti estestvennykh i tekhnicheskikh nauk: sbornik nauchnykh statei IV Nauchno-prakticheskoi mezhdunarodnoi konferentsii (shkoly-seminara) molodykh uchenykh (Tol'iatti, 23–25 aprelia 2018 g.): v 2 ch*. Tol'iatti, Izd-vo Kachalin Aleksandr Vasil'evich, 2018. Part 2. Pp. 162-168.

2. Krasotkin M. A., Shibaikin S. D. Algoritmy bor'by s moshennichestvom v telekommunikatsionnykh sistemakh [Telecommunication fraud fighting algorithms]. *Materialy XXII Nauchno-prakticheskoi konferentsii molodykh uchenykh, aspirantov i studentov natsional'nogo issledovatel'skogo Mordovskogo gosudarstvennogo universiteta im. N. P. Ogareva (Saransk, 25 sentiabria–01 oktiabria 2018 g.): sbornik trudov: v 3 ch*. Saransk, Izd-vo Mordov. gos. un-ta im. N. P. Ogareva, 2019. Part 1. Pp. 148-151.

3. Ladanova E. O., Nikulin V. V. K voprosu ob informatsionnoi bezopasnosti pri analize tekstovykh soobshchenii [On the issue of information security in analysis of text messages]. *Voprosy informatsionnoi bezopasnosti: materialy II Mezhregional'nogo vebinara (Saransk–Elets, 21 fevralia 2018 g.).* Saransk, Izd-vo Mordov. gos. un-ta im. N. P. Ogareva, 2018. Pp. 59-63.

4. Iamashkin S. A. *Metodicheskoe i algoritmicheskoe obespechenie protsessa analiza struktury zemel' na baze dannykh distantsionnogo zondirovaniia. Dissertatsiia ... kand. tekhn. nauk* [Methodological and algorithmic support of land structure analysis based on remote sensing data: Diss. ... Cand. Tech.Sci.]. Saransk, 2016. 186 p.

The article submitted to the editors 18.09.2019

INFORMATION ABOUT THE AUTHORS

Shibaikin Sergei Dmitrievich – Russia, 430005, Saransk; National Research Ogarev Mordovia State University; Candidate of Technical Sciences; Assistant Professor of the Department of Infocommunication Technologies and Communication Systems; shibaikinsd@mail.ru.

Nikulin Vladimir Valer'evich – Russia, 430005, Saransk; National Research Ogarev Mordovia State University; Candidate of Technical Sciences, Assistant Professor; Head of the Department of Infocommunication Technologies and Communication Systems; nikulinvv@mail.ru.

Abbakumov Andrei Aleksandrovich – Russia, 430005, Saransk; National Research Ogarev Mordovia State University; Candidate of Technical Sciences; Assistant Professor of the Department of Automated Information Processing Systems and Management; abbakumov_aa@mail.ru.