Аутентификация пользователя по динамике подписи на основе наивного классификатора Байеса

Гураков М.А., Кривоносов Е.О.
Научный руководитель: доцент каф. КИБЭВС, ТУСУР, к.т.н., доцент Костюченко Е.Ю.

Введение

В настоящее время актуальной задачей остаётся быстрая и простая аутентификация пользователей. Проблемой является то, что быстрая и простая аутентификация – это также ненадёжная аутентификация, в процессе которой велика вероятность как отказа в доступе правомочному пользователю, так и допуска к системе неправомочного. В попытке защитить информационные ресурсы организации устраивают сложные комбинированные системы аутентификации, на что тратятся большие денежные средства, также сотрудникам необходимо запоминать и выполнять некоторую сложную последовательность операций по несколько раз в день. Часто сотрудники в результате незначительных ошибок в процессе аутентификации оказываются не допущены к рабочему месту. Если же ставить системы проще, повышается вероятность доступа внутрь организации злоумышленника, что в потенциале может нанести организации непоправимый ущерб. Сейчас выбор системы аутентификации – это баланс между удобством пользователей и защитой от несанкционированного доступа.

Основная часть

Цель данного проекта заключается в достижении меньшего значения ошибок второго рода при незначительном росте ошибок первого рода разрабатываемой системы аутентификации в сравнении с системами аутентификации на базе нейронных сетей. Идея состоит в совмещении работы алгоритма аутентификации нейронной сети и классификатора Байеса. В рамках проекта аутентификация происходит по подписи пользователя. В среде Matlab R2010a был реализован наивный классификатор Байеса.

Наивный классификатор Байеса служит для распознавания принадлежности объекта некоторому классу, в данном случае для распознавания принадлежности подписи классу подписей отдельного пользователя. Применение классификатора состоит из двух частей – обучения и распознавания.

Алгоритм обучения следующий:

1. Считываются параметры подписей.
2. Область значений параметров разбивается на n-е количество интервалов, количество попаданий в определённый интервал суммируется.
3. Высчитывается $log\frac{P_{ni}(S)}{P_{ni}(\neg S)}$, где $P_{ni}(S)$ – вероятность попадания в n-й интервал для i-го пользователя, а $P_{ni}(\neg S)$ – вероятность попадания в n-й интервал для всех остальных пользователей кроме i-го.

Алгоритм распознавания следующий:

1. Считываются параметры подписи.
2. Область значений параметров разбивается на n-е количество интервалов.
3. Для каждого пользователя высчитывается $\sum_{i=1}^{n}\sum_{j=1}^{m} log\frac{P_i(S)}{P_i(\neg S)}$, где n – количество интервалов, m – количество попаданий в интервал, $P_i(S)$ – вероятность попадания в i-й интервал для конкретного пользователя, $P_i(\neg S)$ – вероятность попадания в i-й интервал для всех пользователей кроме конкретного.

Был произведён комплекс расчётов для определения количества интервалов, при котором обеспечивается минимальное среднее количество ошибок. Расчёты проводились для количества интервалов в диапазоне [2;20] с шагом 1 и [21;25] с шагом 2 (с целью чётко показать характер зависимости), для каждого заданного количества интервалов программа выполнялась 20 раз. Результаты приведены в таблице 1.

Таблица 1 – Результаты тестирования программы на разном количестве интервалов

Количество интервалов	2	3	4	5	6	7	8	9	10	11
Среднее арифметическое ошибки	11,9	12,75	16,9	19,25	24	27,85	27,8	32,45	38,85	45,15
Количество интервалов	12	13	14	15	16	17	18	19	20	21
Среднее арифметическое ошибки	50,35	55,95	60,4	63,25	69,15	74,6	80,8	83,6	90,6	97,8
Количество интервалов	23	25								
Среднее арифметическое ошибки	106,55	126								

На основе полученной зависимости среднего арифметического ошибки от количества интервалов было найдено уравнение регрессии вида $f(x) = 0{,}08253 \cdot x^2 - 2{,}699 \cdot x + 4{,}42$, при нахождении использовался метод наименьших квадратов. График уравнения представлен на графике 1.

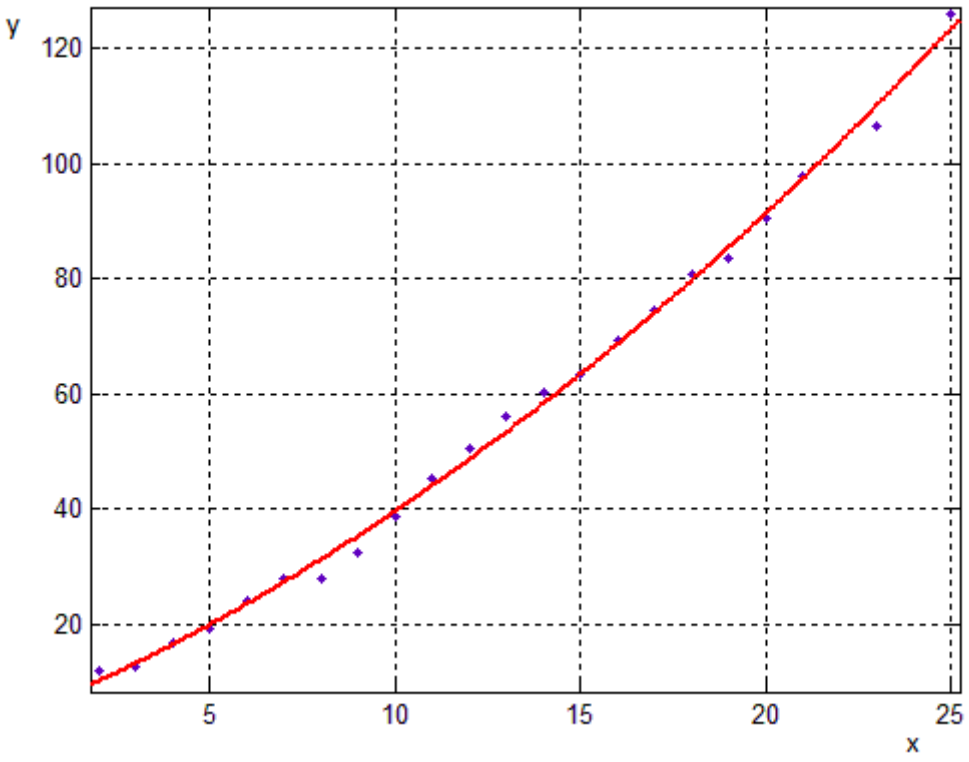

График 1 – График уравнения регрессии на фоне зависимости среднего арифметического ошибки (y) от количества интервалов (x)

Из графика следует, что разбиение области значений на 2 интервала является наилучшим решением. Далее ошибка возрастает в соответствии с найденной зависимостью.

Далее было проведено тестирование. В процессе тестирования программы на 80 процентах сформированной базы подписей (более 1200 подписей, 14 пользователей) производилось обучение, затем на оставшихся 20 процентах производился процесс

распознавания. Каждый раз обучающая и распознаваемая выборки формировались случайным образом. Программа была выполнена 100 раз, область значений параметров разбивалась на 2 интервала. Результаты представлены в таблице 2.

Таблица 2 – Результаты работы программы при количестве интервалов, равном 2

Минимальное количество ошибок, %	Максимальное количество ошибок, %	Среднее выборочное, %	Среднее квадратичное отклонение, %
1,62	10,16	4,78	1,43

Таким образом, средний процент отторжения санкционированных пользователей составляет 4,78%.

Была создана программа, реализующая алгоритм аутентификации на базе нейронной сети, и начаты работы по согласованию работы классификатора Байеса и нейронной сети. Нейронная сеть обучается также на 80 процентах базы подписей, однако выборка формируется внутренними процессами программной среды. Данная выборка извлекается и используется для обучения классификатора Байеса, что обеспечивает равенство условий его обучения с обучением нейронной сети, затем каждый из способов аутентификации применяется на оставшиеся 20 процентов подписей, и результаты сравниваются. На основе получаемых результатов разрабатывается подход к согласованию работы классификатора Байеса и нейронной сети.

Заключение

В ходе проделанной работы были получены следующие результаты:

1. Реализован наивный классификатор Байеса, установлен средний процент отторжения классификатором санкционированных пользователей.
2. Рассчитаны значения количества ошибок при разных делениях области значения параметров и обоснован выбор количества интервалов.
3. Начаты работы с нейронной сетью, сравнения результатов работы нейронной сети и классификатора Байеса на одинаковых выборках.

Следующим этапом исследования будет продолжение формирования базы подписей и согласования работы нейронной сети и классификатора Байеса, а также реализация собственного алгоритма нормализации, после которого ожидается снижение ошибок распознавания санкционированных пользователей.

Литература

1. Субботин С. В., Большаков Д. Ю. Применение байесовского классификатора для распознавания классов целей. // «Журнал Радиоэлектроники», 2006, № 4
2. McCallum, A. and Nigam K. «A Comparison of Event Models for Naive Bayes Text Classification». In AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48. Technical Report WS-98-05. AAAI Press. 1998.
3. Т.Ю. Дорошенко, Е.Ю. Костюченко. Система аутентификации на основе динамики рукописной подписи. // Доклады ТУСУРа, № 2 (32), июнь 2014, стр. 219-223