

# **АНАЛИЗ СТРАНИЦ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ “ВКОНТАКТЕ” С ЦЕЛЮ ВЫЯВЛЕНИЯ СОТРУДНИКОВ ЗАДАННОЙ КОМПАНИИ<sup>1</sup>**

Шиндарев Н.А., студент Санкт-Петербургского государственного университета, nshindarev@gmail.com

Абрамов М.В., младший научный сотрудник лаборатории ТиМПИ СПИИРАН; старший преподаватель СПбГУ, mva16@list.ru

Тулупьев А.Л., д.ф.-м.н., доц., зав.лаб. ТиМПИ СПИИРАН; проф. Санкт-Петербургского государственного университета, alexander.tulupyev@gmail.com

## **Аннотация**

В докладе приводится описание моделей и алгоритмов реализованного программного модуля, анализирующего страницы пользователей социальной сети “ВКонтакте” на предмет аффилированности с некоторой заданной компанией. Программный модуль осуществляет сбор обучающей и тестовой выборки, на основе которых строится дерево принятия решений. Результатом прохода по дереву решений является утверждение о принадлежности страницы в социальной сети “ВКонтакте” одному из сотрудников компании.

## **Введение**

В данной статье представлены основные этапы построения собственного классификатора для анализируемой компании [1,2,3]. На основе этого классификатора в дальнейшем будет осуществляться сбор пользовательских страниц социальной сети «ВКонтакте» [5,6], которые принадлежат сотрудникам организации. Представленная работа является частью глобального проекта, конечной целью которого является автоматизация этапов составления профиля уязвимостей сотрудников компании. Итоговое программное решение при внедрении в существующий программный комплекс по анализу информационных систем на предмет уязвимостей к социоинженерным атакам должно

---

<sup>1</sup> Работы выполнялись в рамках проекта по государственному заданию СПИИРАН № 0073-2014-0002.

повысить точность результирующей оценки информационной системы в силу того, что социальные сети зачастую содержат в себе личную информацию о пользователе, которая не учитывается при социологических опросах пользователей.

## Формализация задачи

В рамках данной работы описываются детали выполнения той части задач, которая относится к построению классификатора. Поставленная задача выявления пользовательских страниц сводится к задаче бинарной классификации [9] при следующей формализации:

Пусть  $X$  — множество страниц пользователей социальной сети vk.com, а  $Y$  — множество наименований классов (в данном случае  $|Y| = 2$ , т.к. мы имеем в результате всего 2 класса: сотрудники и не сотрудники). Существует целевая зависимость:

$$y^* : X \rightarrow Y$$

При этом значения для неё известны только на конечном числе объектов обучающей выборки:

$$X^m = \{ (x_1, y_1), (x_2, y_2), \dots, (x_m, y_m) \}$$

Тогда задача сводится к построению алгоритма

$$a : X \rightarrow Y$$

способного классифицировать любой  $x \in X$ . Классификация при  $|Y| = 2$  называется бинарной. [5, 19] (ссылки в дипломе)

## Описание классификатора

Для решения задачи классификации в рамках данной работы применяется структура дерева принятия решений. Такой выбор объясняется тем, что в ряде источников рекомендуется использовать данную структуру в случае, если выполняются следующие условия:

Среди прочих методов Data Mining, метод дерева принятия решений имеет несколько достоинств:

- прост в понимании и интерпретации;
- данные не требуют подготовки: прочие техники требуют нормализации данных, добавления фиктивных переменных, а также удаления пропущенных данных;
- способен работать как с категориальными, так и с интервальными переменными;
- использует модель «белого ящика»: если определенная ситуация наблюдается в модели, то её можно объяснить при помощи булевой логики;

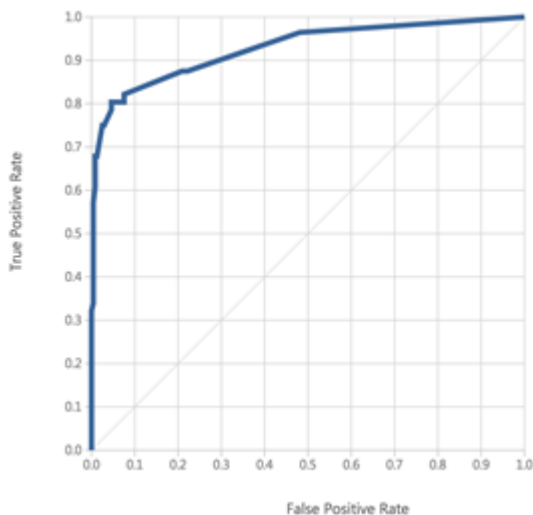
- позволяет оценить модель при помощи статистических тестов. Это дает возможность оценить надежность модели;
- метод хорошо работает даже в том случае, если были нарушены первоначальные предположения, включенные в модель;
- позволяет работать с большим объемом информации без специальных подготовительных процедур. Данный метод не требует специального оборудования для работы с большими базами данных.

Также стоит отметить, что в данная структура была применена в исследовании [8], в котором анализировался онлайн-след сотрудников компании. Структура показала хорошие результаты, качество результирующего дерева было оценено на основе показателя  $f_1 - score$ [10]. Как итог, дерево принятия решений показал рейтинг  $f_1 - score = 0.65$ , в то время как альтернативная структура на основе “случайного леса” выдала показатели:  $precision = 0.67$ ,  $recall = 0.08$  и результирующее значение  $f_1 - score = 0.14$ , что является недопустимым значением, поскольку в контексте поставленной задачи в первую очередь нам важна полнота полученных результатов.

Программное решение для поставленных задач разработано на языке C# для платформы .NET с учётом возможности дальнейшей интеграции программы в существующий комплекс в виде динамически подключаемой библиотеки. Для обучения и дальнейшей эксплуатации деревьев решений используется фреймворк Accord .NET [12].

## Анализ параметров и оценка полученного классификатора

Для оценки качества классификатора используется ROC-кривая, которая при варьировании порога решающего правила показывает то, как зависит recall от FPR (False Positive Rate). Ниже представлена ROC-кривая для анализа эффективности полученного классификатора на примере одной Российской ИТ-компании с заявленной численностью штата в 1200 сотрудников:



**Рис.1.** Пример ROC-кривой

Для получения числовой характеристики ROC-кривой используется площадь под графиком AUC (Area Under Curve). При  $AUC = 1$  считается, что дерево выдает случайные значения, и, соответственно, чем ближе значение к 1, тем лучше получился классификатор. В примере на рис.1 получившийся показатель  $AUC = 0.928$ , что является очень хорошим результатом.

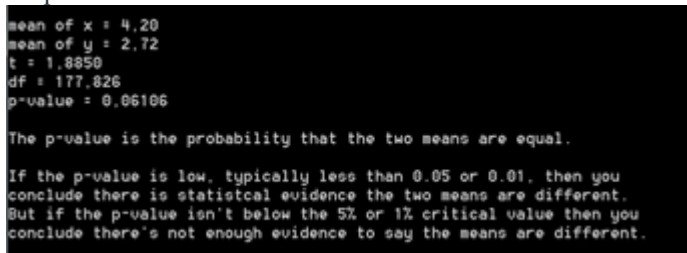
Для принятия решений в ходе исследования было выявлено 5 основных параметров, каждый из которых фиксирует какую-либо из поведенческих тенденций, свойственных сотрудникам компаний:

- 1) наличие названия компании в графе «Карьера»;
- 2) упоминание имени сотрудника на стене официальной группы компании;
- 3) результат анализа топологии сети для данной страницы;
- 4) проверка наличия данной страницы в списке подписок компании;
- 5) счётчик отметок «Мне нравится», оставленных данным пользователем на стене группы компании.

Чтобы повысить качество результирующего дерева необходимо строить собственный бинарный классификатор для каждой анализируемой компании. Это возможно только при полной автоматизации сбора и анализа целевой переменной для обучающей выборки. Для этого было решено добавлять в обучающую выборку страницы с указанным местом

работы в анализируемой компании в качестве примеров со значением целевого параметра = 1. На основе списка их друзей обучающая выборка дополнялась страницами пользователей, у которых указано текущее место работы в другой компании. Значение целевого параметра у этих страниц = 0.

В результате обучающая выборка была составлена исключительно на основе страниц, в которых пользователи указали текущее место работы. Для обоснования того, что данное допущение не искажает результатов, необходимо проверить гипотезу о равенстве математических ожиданий двух случайных величин: одна фиксирует численное значение, характеризующее какое-либо поведение пользователя в группе для страниц с указанным местом работы, а другая - для страниц без указанного места работы. Для выполнения проверки гипотезы использовался t-test Стьюдента [13,14]. В результате работы данного теста для компаний была получена вероятность равенства математических ожиданий для данных случайных величин. В случае, если вероятность составляет меньше 3%, можно заявлять о том, что математические ожидания не совпадают. Ниже, на рис.2, приведены результаты работы t-теста для компании, для которой на рис.1 представлена кривая ошибок. Результат показал вероятность = 8%, что позволяет нам использовать допущение о выборе страниц обучающей выборки в работе.



```
mean of x = 4.20
mean of y = 2.72
t = 1.8850
df = 177.826
p-value = 0.06106

The p-value is the probability that the two means are equal.

If the p-value is low, typically less than 0.05 or 0.01, then you
conclude there is statistical evidence the two means are different.
But if the p-value isn't below the 5% or 1% critical value then you
conclude there's not enough evidence to say the means are different.
```

**Рис.2.** Пример результата работы t-теста Стьюдента

## **Заключение**

В результате проведённого эксперимента была получена модель, которая достаточно точно идентифицирует страницы аккаунтов сотрудников компании в социальной сети «ВКонтакте» на основе анализа комплекса параметров. Основная возможность увеличения точности идентификации аккаунтов сотрудников организации в социальной сети «ВКонтакте» видится посредством проведения дополнительного, более детального анализа возможных признаков, которые позволяли бы делать вывод о принадлежности профиля сотруднику.

## Литература

1. Abramov M. V., Azarov A. A. Social engineering attack modeling with the use of Bayesian networks //Soft Computing and Measurements (SCM), 2016 XIX IEEE International Conference on. – IEEE, 2016. – p. 58-60.
2. Abramov M.V., Azarov A.A., Tulupyeva T.V., Tulupyevev A.L. Model of Malefactor Profile for Analyzing Information System Personnel Security from Social Engineering Attacks //Information and Control System. 2016. No4. p.77–84
3. Azarov A.A., Tulupyeva T.V., Suvorova A.V., Tulupyevev A.L., Abramov M.V., Usypov R.M. Social Engineering attacks: problem of analisys. — SPb.: Science, 2016.
4. Information security business. Studies of current trends in information security business // Kaspersky Lab. URL: [http://media.kaspersky.com/pdf/IT\\_risk\\_report\\_Russia\\_2014.pdf](http://media.kaspersky.com/pdf/IT_risk_report_Russia_2014.pdf) (date of the application: 30.04.2015)
5. Social networks in Russia, autumn 2016. Numbers, trends, forecasts. – URL: <https://adindex.ru/publication/analitics/100380/2016/12/8/156545.phtml> (online; accessed: 13.04.2017)
6. Social networks in Russia, winter 2015-2016. Numbers, trends, forecasts. – URL: <https://blog.br-analytics.ru/socialnye-seti-v-rossii-zima-2015-2016-cifry-trendy-prognozy/> (online; accessed: 13.04.2017)
7. The losses from cybercrime continue to grow // URL: <http://www8.hp.com/ru/ru/software-solutions/ponemon-cyber-security-report/index.html> (date of the application: 12.04.2017).
8. *Edwards M. et al.* Panning for gold: automatically analysing online social engineering attack surfaces //Computers & Security. – 2016.
9. Wikipedia: Binary classification. URL: [https://en.wikipedia.org/wiki/Binary\\_classification](https://en.wikipedia.org/wiki/Binary_classification)
10. Wikipedia: Precision and recall. URL: [https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall)
11. Wikipedia: Decision Trees. URL: [https://en.wikipedia.org/wiki/Decision\\_tree](https://en.wikipedia.org/wiki/Decision_tree)
12. Сайт проекта accord.net. URL: <http://accord-framework.net/>.
13. Wikipedia: t-критерий Стьюдента. URL: [https://en.wikipedia.org/wiki/Student%27s\\_t-test](https://en.wikipedia.org/wiki/Student%27s_t-test).
14. Тесты - Т тест на C#, november 2015. URL: <https://msdn.microsoft.com/ru-ru/magazine/mt620016.aspx>