element method. *Telecommunications and Radio Engineering*, 2013, vol. 72, pp. 111–123. doi: 10.1615/TelecomRadEng.v72.i2.30
37. Gradstein I.S., Ryjik I.M. *Tablicy integralov* [Tables of integrals]. Moscow, GIFML Publ., 1962. 1100 p.
38. Definition and test methods for the relevant parameters of single-mode fibres. ITU COM 15-273-E. 1996.
39. Dubois F., Emplit Ph., Hugon O. Selective mode excitation in graded-index multimode fiber by a computer-generated optical mask. *Optics Letters*, 1994, vol. 19, pp. 433–435. doi: 10.1364/OL.19.000433
40. Karpeev S.V., Pavelyev V.S., Soifer V.A., Doskolovich L.L., Duparre M., Luedge B. Mode multiplexing by diffractive optical elements in optical telecommunication. *Proc. SPIE 5480*, 2004, pp. 163–165. doi:10.1117/12.558775
41. Noordegraaf D., Skovgaard P.M., Nielsen M.D., Bland-Hawthorn J. Efficient multi-mode to single-mode coupling in a photonic lantern. *Optics Express*, 2009, vol. 17, pp. 1988–1994. doi: 10.1364/OE.17.001988
42. Leon-Saval S.G., Argyros A., Bland-Hawthorn J. Photonic lanterns: a study of light propagation in multimode to single-mode converters. *Optics Express*, 2010, vol. 18, pp. 8430–8439. doi: 10.1364/OE.18.008430

Received 25.08.2016

УДК 004.4(075)

КЛАССИФИКАЦИЯ БОЛЬШИХ ДАННЫХ ТЕЛЕКОММУНИКАЦИОННОЙ КОМПАНИИ С ПОМОЩЬЮ ТЕХНОЛОГИИ DATA MINING

Самаркин М.Е., Тарасов В.Н.
Поволжский государственный университет телекоммуникаций и информатики, Самара, РФ
E-mail: m.samarkin@psuti.ru, tarasov-vn@psuti.ru

С помощью технологии data mining анализируется информация о клиентах телекоммуникационной компании. К методу кластерного анализа k-средних применен нестандартный метод инициализации центроидов k-means++. Разработаны программа на языке C#, которая преобразует исходные данные компании в формат Comma-Separated Values (CSV), и программа на языке Python для кластеризации данных с подключенной библиотекой sklearn. Представлен обученный на больших данных классификатор на основе метода опорных векторов SVM.

Ключевые слова: кластерный анализ, метод k-средних, метод инициализации центроидов k-means++, C#, Python, oneClassSVM.

Введение

В телекоммуникационных компаниях (ТК) с использованием баз данных (БД) ежедневно собираются большие объемы информации. Из такого объема данных необходимо извлечь ценную информацию, касающуюся работы компании и ее клиентов. Для этой цели хорошо подходит технология DATA MINING – как мультидисциплинарная область, возникшая и развивающаяся на базе таких наук, как прикладная статистика, классификация и кластеризация, распознавание образов, искусственный интеллект, теория баз данных, и др. Поэтому здесь хорошо просматривается интеграция теории прикладной статистики анализа данных с очисткой данных, обучением и визуализацией результатов.

Применительно к обработке информации ТК главная проблема состоит в том, что зачастую заранее неизвестно, как группировать эти данные, как находить связи между событиями и т.д. Для частичного решения данной проблемы используются методы классификации и кластеризации DATA MINING. При помощи них можно объединять данные по любым выбранным признакам. Например, находить взаимосвязь события от нагрузки в сети или дня недели и т.п.

Постановка задачи

В качестве примера накопления больших объемов данных рассмотрим ТК, которые каждый день собирают информацию о трафике пользователей. Требуется проанализировать весь объем информации, полученный за один месяц работы, и выявить аномалии, нестандартное поведение клиентов и соотношение объема трафика ко дню недели и типу дня (рабочий, выходной).

Компанией «СамараСвязьИнформ» авторам был предоставлен массив данных (135 Гб в формате txt) о трафике, который клиенты генерируют ежедневно. Он содержит в числе прочих следующую информацию: ID аккаунта, IP адрес источник, IP адрес получатель, размер пакета данных, номер порта источника, номера порта получателя, день и время события в unix формате. В общей сложности это примерно 45 млн строк вида, показанного на рис. 1.

timestamp	1456634876
account_id	0
source	192.168.101.3
destination	178.218.82.79
t_class	2210
packets	1
bytes	258
sport	53
dport	57577
date	Sun Feb 28 08:47:56 2016

Рис. 1. Пример строки предоставленных данных

Классическая задача DATA MINING состоит в том, чтобы из очень большого объема данных извлечь самую нужную и ценную информацию. Провести анализ таких данных вручную невозможно, поэтому было принято решение преобразовать данные в нужный формат и применить к решению поставленной задачи статистические методы кластеризации и классификации.

Кроме того, для анализа таких объемов информации также невозможно применить готовые статистические пакеты. Известный метод классификации на основе корреляционного анализа также здесь не может быть применен вследствие того, что получится корреляционная матрица очень большого порядка. Ниже приведена последовательность действий, необходимых для получения реальных результатов.

Первичная обработка данных

Для выявления соотношения объема трафика и активности клиента в целом решено преобразовать данные в следующий формат: ID аккаунта, объем трафика за сутки в гигабайтах, день недели, порядковый номер дня, тип дня (выходной – 0, рабочий – 1), число запросов за сутки – всего шесть упорядоченных признаков каждого вектора данных.

Для такого преобразования была написана программа на языке C#, которая открывала по очереди txt файлы на чтение, складывала трафик за сутки каждого клиента и считывала количество обращений этого клиента к серверу. После преобразований было получено 60473 строк информации, и если поделить это число на количество дней в месяце, то мы получим количество активных абонентов компании. Каждая строка – это информация о клиенте за день. На рис. 2 приведен пример данных для пяти клиентов компании с шестью упорядоченными признаками.

```
0,30.0108,1,1,1,13883303
1,0.007,1,1,1,4573
10,0.0341,1,1,1,3333
29,0.7152,1,1,1,38790
66,0.0716,1,1,1,4737
```

Рис. 2. Пример данных после преобразования

Описание программы преобразования данных к кластеризации

Алгоритм работы этой программы разбивается на несколько цикличных этапов:

1. Открытие текстовых файлов с информацией, пока не прошли по всем файлам.
2. Построчное считывание информации, пока не конец файла.
3. Разбор строки на составляющие: id аккаунта, размер пакета и т.д.
4. Суммирование объема трафика.
5. Подсчет количества обращений к серверу.
6. Преобразование данных в формат CSV и сохранение полученных данных в файл.

После первичной обработки требуется разбить эти данные на группы и проанализировать полученный результат. Поскольку при этом априори предполагается, что среди клиентов компании по двум признакам (второй и шестой) могут быть нестандартные, для этого был выбран метод кластеризации k-средних.

Кластеризация данных

Точность разбиения полученных векторов по кластерам зависит от изначально выбранных центроидов. Нами выбран метод инициализации центроидов под названием k-means++. Это улучшенная версия алгоритма кластеризации k-средних. Суть улучшения заключается в нахождении более «хороших» начальных значений центроидов кластеров. Алгоритм метода включает пять основных шагов.

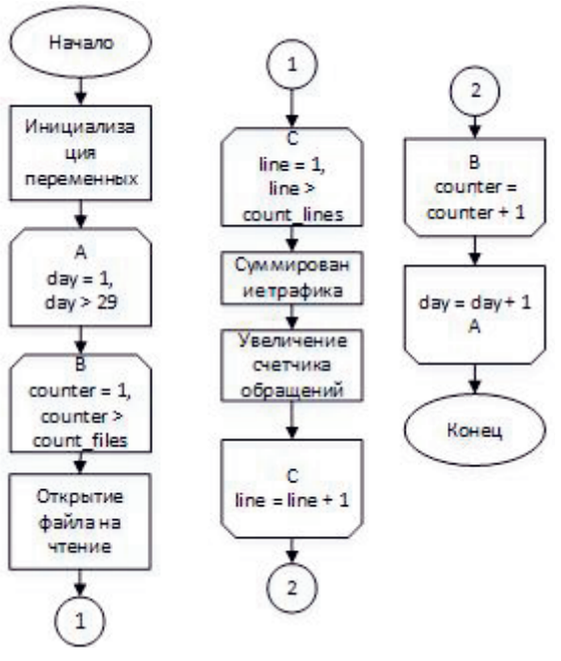

Рис.3. Схема алгоритма работы программы

Инициализация.

1. Выбрать первый центроид случайным образом (среди всех точек).

2. Для каждой точки найти значение квадрата расстояния до ближайшего центроида (из тех, которые уже выбраны) $(dx)^2$.

3. Выбрать из этих точек следующий центроид так, чтобы вероятность выбора точки была пропорциональна вычисленному для нее квадрату расстояния.

4. Это можно сделать следующим образом: на шаге 2, параллельно с расчетом $(dx)^2$, нужно подсчитывать сумму $Sum(dx^2)$. После накопления данной суммы найти значение Rnd=random(0.0,1.0)*Sum. Генератор Rnd случайным образом укажет на число из интервала [0; Sum), и остается только определить, какой точке это соответствует. Для этого нужно снова начать подсчитывать сумму $S(dx^2)$ до тех пор, пока сумма не превысит Rnd. Как только это случится, суммирование останавливается и мы можем взять текущую точку в качестве центроида. При выборе каждого следующего центроида не надо следить за тем, чтобы он не совпал с одной из уже выбранных в качестве центроидов, так как вероятность повторного выбора точки равна нулю.

5. Повторять шаги 2 и 3 до тех пор, пока не будут найдены все необходимые центроиды.

Далее выполняется основной алгоритм k-средних: этот популярный метод кластеризации был предложен полвека назад Г. Штейнгаузом и С. Ллойдом. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров: $V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \to \min$, где k – число кластеров, S_i – полученные кластеры, $i = 1; 2 \ldots k$ и μ_i – центры масс векторов x_j, принадлежащих S_i.

Анализ полученных результатов

Для проведения кластерного анализа была написана программа на языке Python с подключенной библиотекой sklearn, которая содержит в себе реализации большинства математических методов интеллектуального анализа данных. В укрупненном виде схема алгоритма программы приведена на рис. 4.

Рис. 4. Схема алгоритма программы кластеризации

Применив совмещенный метод k-средних с методом инициализации центроидов k-means++, и разбив данные на два кластера, получим следующие результаты: 1-ый кластер включает 60427 записей; 2-ой кластер – 46 записей.

На рис. 5 можно видеть графическое распределение векторов (записей) по кластерам. Здесь по горизонтальной оси X указано количество ответов от сервера, а по вертикальной оси Y – account id. Белыми крестами отображены центроиды кластеров.

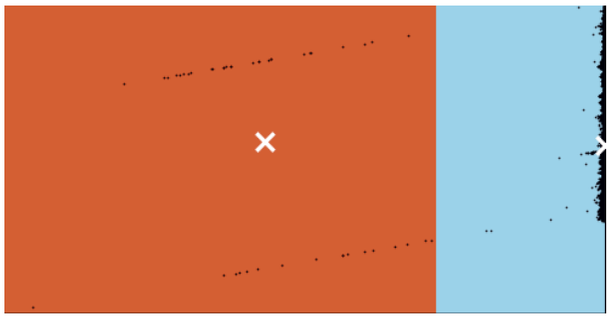

Рис. 5. Кластеры, полученные методом k-средних

Как можно интерпретировать полученные данные? Первый кластер (справа) содержит информацию о стандартных записях, которые не выделяются из общей массы. Это стандартное поведение клиентов в течение месяца. Второй кластер (слева) содержит информацию о выделившихся из массы нестандартных записях. Это «особое», отличающееся от нормального, поведение клиентов: скорее всего, на стороне клиентов с такими признаками можно обнаружить вирусную активность либо «плохое» программное обеспечение, вызывающие очень большое число обращений к серверу ТК.

Построение классификатора и его обучение

Для того, чтобы программное обеспечение в дальнейшем само автоматически определяло принадлежность той или иной записи к выделенным кластерам, необходимо применить метод классификации. Для этого рассмотрим метод опорных векторов (англ. SVM, support vector machine) – набор схожих алгоритмов обучения с учителем, использующихся для задач классификации и регрессионного анализа. Он принадлежит к семейству линейных классификаторов, может также рассматриваться как специальный случай регуляризации по А.Н. Тихонову. Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора между гиперплоскостями, поэтому метод также известен как метод классификатора с максимальным зазором.

Основная идея метода – перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельные гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей разные классы. Разделяющей гиперплоскостью будет гиперплоскость, которая максимизирует расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Для применения метода классификации к данным каждая запись должна быть вручную помечена принадлежностью к конкретному классу – после чего данные выглядят, как это показано на рис. 6. Вводится новое поле target, которое хранит в себе информацию о принадлежности каждой записи к тому или иному классу.

account_id	0
traffic	30.0108
count	13883303
dayOfWeek	1
n_day	1
type_day	1
month	2
target	1

Рис. 6. Вид данных

Здесь поле target – это классификатор: 1 – относит данные к первому кластеру; 0 – ко второму.

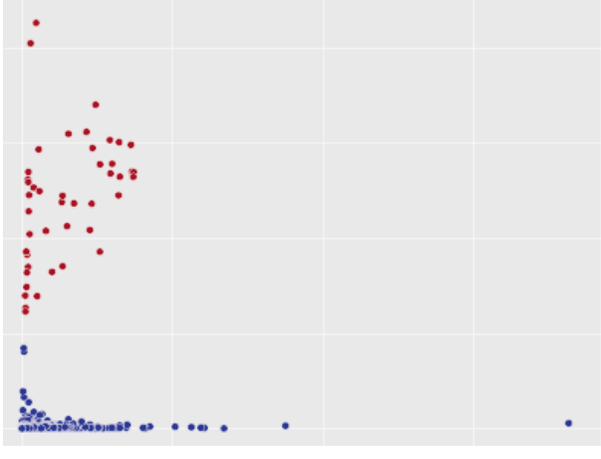

Рис. 7. Графический вид классифицированных данных

На рис. 7 отображены классифицированные данные по полю target: на экране красным цветом обозначены векторы, принадлежащие к кластеру 1 (верхнее семейство точек), синим цветом – к кластеру 2 (нижнее семейство точек). Программа написана на языке Python, также с использовани-

ем библиотеки sklearn, которая классифицирует новые данные при помощи метода опорных векторов и выводит информацию о принадлежности этих данных к тому или иному классу. Укрупненная схема алгоритма программы приведена на рис. 8.

Для примера проверим принадлежность к кластерам следующие записи (размер трафика в гигабайтах, число ответов от сервера, день недели, день месяца, тип дня, месяц):

1. 30.0108,13883303,1,1,2,1
2. 0.007,4573,1,1,2,1,0.

Программа дает ответ: первая запись принадлежит к первому кластеру, вторая – ко второму кластеру, что полностью подтверждает правильность работы классификатора.

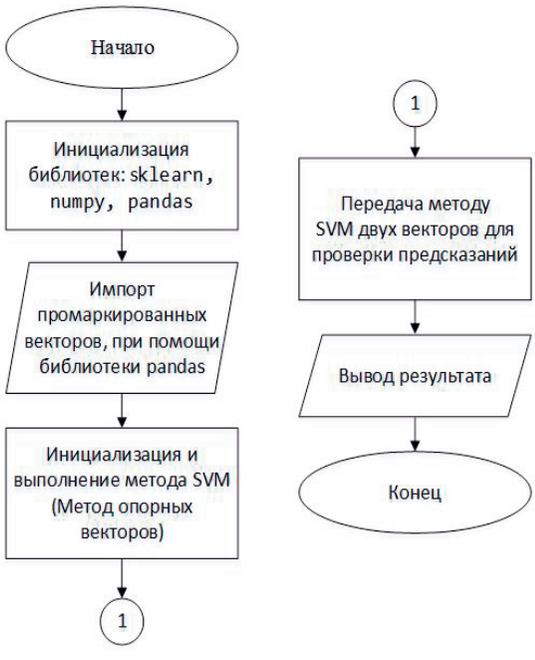

Рис.8. Схема алгоритма программы классификации данных

Заключение

Проведен полный спектр анализа данных ТК, предоставленных за февраль, которые содержат служебную информацию об IP-потоках. Для проведения полного анализа написана программа на языке C#, агрегирующая данные по дням, – таким образом были выделены признаки, в соответствии с которыми в дальнейшем будет осуществляться классификация по известным методам.

Проведен кластерный анализ при помощи авторской программы на языке Python, выявлены два кластера: среднестатистические клиенты и «особые». Информация отображена на графиках. По этим данным было проведено обучение выборки большого объема (60427 векторов) с введением поля «target».

Применен метод опорных векторов для классификации новых данных, поступающих в реальном времени. Обученный классификатор может быть применен для анализа новых данных ТК.

Литература

1. Метод инициализации центроидов K-means++ // URL: https://ru.wikipedia.org/wiki/K-means++ (д.о. 10.03.2016).
2. Метод кластеризации K-средних // URL: https://ru.wikipedia.org/wiki/K-means (д.о. 02.02.2016).
3. Метод опорных векторов // URL: http://www.machinelearning.ru/wiki/index.php?title=SVM (д.о. 11.01.2016).
4. Машинное обучение (Воронцов К.В., курс лекций) // URL: http://www. machinelearning.ru/wiki/index.php?title(д.о. 11.01.2016).
5. Richert W., Coelho L.P. Building Machine Learning Systems with Python. 2013. – 290 p.
6. Leskovec J., Rajaraman A., Ullman J.D. Mining of Massive Datasets. 2014. – 476 p.
7. Flach P. Machine Learning. The Art and Science of Algorithms that Make Sense of Data. 2012. – 409 p.
8. McKinney W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. 2012. – 400 p.
9. Тарасов В.Н., Самаркин М.Е. Подходы к кластеризации больших данных по событиям в сети // Тезисы XXIII РНТК ПГУТИ, 2016. – С. 253.
10. Тарасов В.Н. Об одном из способов повышения надежности классификационного анализа // Интеллект. Инновации. Инвестиции. №4, 2014. – С. 107-111.

Получено 08.06.2016

Самаркин Михаил Евгеньевич, магистр Кафедры программного обеспечения и управления в технических системах (ПОУТС) Поволжского государственного университета телекоммуникаций и информатики (ПГУТИ). Тел. (8-846) 228-00-13; E-mail: m.samarkin@psuti.ru

Тарасов Вениамин Николаевич, д.т.н., профессор, заведующий Кафедрой ПОУТС ПГУТИ. Тел. (8-846) 228-00-13; E-mail: tarasov-vn@psuti.ru

TELECOMMUNICATION COMPANY BIG DATA CLASSIFICATION BY DATA MINING TECHNIQUE

Samarkin M.E., Tarasov V.N.
Povolzhsky State University of Telecommunication and Informatics, Samara, Russian Federation
E-mail: m.samarkin@psuti.ru, tarasov-vn@psuti.ru

This work deals with analysis of telecommunication company customer information by data mining technique. For this purpose, we used cluster analysis method of k-means and support vector machines. The first was applied with non-conventional k-means++ centroid initialization method. Special software on C# language was developed that transforms initial data of telecommunication company to comma-separated values. We designed software on Python language with library sklearn for data clasterization. Data processing was performed, and customers with "aberrant behavior" were detected. Telecommunication company should make decision for those customers by itself. We developed classifier learned on big data based on support vector machines to classify new data of company.

Keywords: cluster analysis, k-means method, k-means++ centroid initialization method, C #, Python, oneClassSVM

DOI: 10.18469/ikt.2016.14.3.05

Samarkin Michael Evgenevich, Povolzhsky State University of Telecommunications and Informatics, 23 Lev Tolstoy str., Samara 443010, Russian Federation; magistrand of the Department of Software and Management in Technical Systems. Tel.: +78462280013. E-mail: m.samarkin@psuti.ru.

Tarasov Veniamin Nikolayevich, Povolzhsky State University of Telecommunications and Informatics, 23 Lev Tolstoy str., Samara 443010, Russian Federation; the Head of Department of Software and Management in Technical Systems, Doctor of Technical Science, Professor. Tel.: +78462280013; E-mail: tarasov-vn@psuti.ru.

References

1. Metody initsializatsii zentroidov K-means++ [Method initialization centroid. K-means++]. Available at: https://ru.wikipedia.org/wiki/K-means++ (accessed 10.03.2016),
2. Metod klasterizatsii k-srednich [Clustering method K-means]. Available at: https://ru.wikipedia.org/wiki/K-means (accessed 11.01.2016),
3. Metod opornych vectorov [Method of support vector machine]. Available at: http://www.machinelearning.ru/wiki/index.php?title=SVM (accessed 11.01.2016).
4. Mashinnoe obuchenie (kurs lekcii, K.V. Voroncov) [Machine learning. Lecture course K.V. Voroncov]. Available at: http://www. machinelearning.ru/wiki/index.php?title (accessed 11.01.2016).
5. Richert W., Coelho L. *Building Machine Learning Systems with Python*, Packt Publishing, 2013. 290 p.
6. Leskovec J., Rajaraman A., Ullman J. *Mining of Massive Datasets*. Cambridge University Press, 2014. 476 p.
7. Flach P. *Machine Learning. The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012. 409 p.
8. McKinney W. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, 2012. 400 p.
9. Samarkin M.E., Tarasov V.N. Podchody k klasterizatsii bolshich dannych po sobytiyam v seti [Approaches to the clustering big data network events]. *Materialy XXIII Rossiiskoi nauchnoi konferenzii PPS, NS i aspirantov*, PGUTI, 2016, p. 253.
10. Tarasov V.N., Ob odnov is sposobov povyshenia nadezhnosti klassifikacionnogo analiza [On one of the ways to improve the reliability of the classification analysis]. *Intelekt. Innovacii. Investicii*. 2014, no.4, pp.107-111.

Received 08.06.2016