Санкт-Петербургский государственный университет

Кафедра Системного Программирования

Кравченко Евгений Артурович

Применение метода опорных векторов в задаче предсказания оттока клиентов оператора мобильной связи

Курсовая работа

Научный руководитель:
д. ф.-м. н., профессор Терехов А. Н.

Санкт-Петербург
2018

Оглавление

Введение

Ежегодно провайдеры телекоммуникационных услуг терпят убытки из-за оттока абонентов. Сегодня существует большое количество различных поставщиков телекоммуникационных услуг, при этом процесс смены провайдера с каждым годом становится все проще. К тому же после вступления в силу закона о MNP(Mobile Number Portability), появилась возможность перенести старый номер при смене оператора мобильной связи. По этим причинам годовой отток телекоммуникационных компаний может достигать 50% (рис 1).

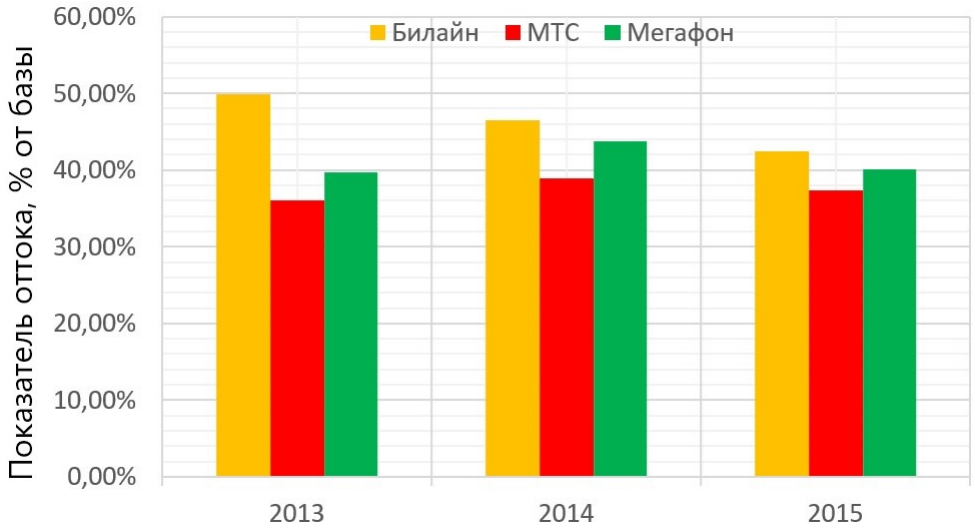

Рис. 1: Отток клиентов мобильных операторов в России[1]

В связи с тем, что удержание текущих клиентов обходится в несколько раз дешевле, чем привлечение новых, точные предсказания оттока абонентов позволяют телекоммуникационным компаниям применять различные стратегии для удержания клиентов и тем самым экономить большое количество средств.

Задача предсказания оттока является классической задачей бинарной классификации. Всех абонентов необходимо разделить на две группы: те, кто скорее всего продолжит пользоваться услугами компании и те, кто, вероятнее всего, уйдут. В решении такого рода задач наиболее хорошо себя зарекомендовали различные методы машинного обучения,

[1]Согласно отчету о рынке мобильной связи в 2015 году, https://static.beeline.ru/upload/images/Beeline_MKT_Overview_2015F.pdf

такие как решающие деревья, нейронные сети, логистическая регрессия, метод опорных векторов и другие. Также обширно используются разнообразные способы построения композиций простых алгоритмов. В данной работе будут рассмотрены возможности метода опорных векторов и его модификаций в решении задачи предсказания оттока.

1. Цель работы

Цель данной работы - для предоставленной выборке разработать классификатор, основанный на методе опорных векторов векторов для задачи предсказания оттока абонента оператора мобильной связи и рассмотреть применимость данного метода к этой задаче. Для этого требуется решить следующие задачи:

Задачи:

1. Рассмотреть уже существующие решения данной задачи.
2. Подготовить предоставленные данные.
3. Разработать классификатор, предсказывающий уход абонентов.
4. Оптимизировать классификатор для данной задачи.
5. Рассмотреть возможности ансамблей с использованием SVM.

2. Терминология

- *Образец* - вектор вещественных чисел (характеристик).
- *Выборка* - конечный набор объектов.
- *Тестовая выборка* - выборка, по которой оценивается качество построенной модели (алгоритма).
- *Классификатор* - отображение $X \to \{1, 2, ..., n\}$ из множества образцов в множество классов. Если n = 2, то классификатор называют бинарным.
- *Ансамбль* - композиция нескольких методов машинного обучения.
- *Бэггинг* - это композиция алгоритмов, каждый из которых обучается независимо.
- *SVM* - метод опорных векторов (Support Vector Machine) - методология обучения по прецедентам, предложенная В.Н. Вапником.

3. Оценка точности классификации

Весь набор данных разделим на 4 группы:

- True Positives (TP) - верно определенные в положительный класс (уходящие абоненты).
- False Positives (FP) - ошибочно определенные в положительный класс.
- True Negatives (TN) - верно определенные в отрицательный класс (удержанные абоненты).
- False Negatives (FN) - ошибочно определенные в отрицательный класс.

Определим вспомогательные параметры

$$P = TP + FN, \quad N = TN + FP$$

Для измерения точности бинарного классификатора обычно используются следующие характеристики:

- Точность
$$precision = \frac{TP}{TP + FP}$$
Показывает то, какая доля объектов, определенных классификатором в положительный класс, действительно является положительной.

- Полнота
$$recall = \frac{TP}{P}$$
Показывает то, какую часть объектов положительного класса классификатор определил верно.

Какая из данных характеристик важнее, зависит от конкретной задачи. Следующая функция позволяет придать различный вес точности и полноте в зависимости от параметра β.

- F - мера

  Один из способов получения критерия качества классификатора на основе на точности и полноты.

$$F_\beta = \left(\beta^2 + 1\right) \frac{precision \times recall}{\beta^2\ precision + recall}$$

  (при $0 < \beta < 1$ приоритет отдается точности, при $\beta > 1$ - полноте)

Параметр β зависит только от внешних факторов, которые не рассматриваются в рамках данной работы, поэтому далее будет вычисляться F_1 - мера, то есть среднее гармоническое $precision$ и $recall$.

Для оценки качества бинарной классификации также используется понятие ROC - кривой, которая задается параметрически: $x = \frac{FP}{N}, \quad y = \frac{TP}{P}$. Следующее понятие является количественной характеристикой для ROC - кривой:

- AUC (Area under ROC curve)

$$AUC = \int_0^1 \frac{TP}{P}\ d\frac{FP}{N}$$

  ($0 \leq$ AUC ≤ 1. Если AUC $= 0.5$, то классификатор равен случайному, если AUC < 0.5, то необходимо инвертировать ответы классификатора. Идеальным случаем для классификатора является AUC $= 1$)

Для преобразования вещественного ответа алгоритма (величины, показывающей "степень уверенности", с которой алгоритм отнес данный объект к данному классу) в бинарную метку, используется *порог*: все объекты, для которых результат работы алгоритма больше порога, определяются в положительный класс, остальные - в отрицательный. ROC AUC позволяет оценить модель в целом, не привязываясь к конкретному порогу.

4. Обзор

Данная задача уже рассматривалась во многочисленной литературе. В таблице ниже приведены рассмотренные работы.

Автор	Работа
Kristof Coussement, Dirk Van den Poel	Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques [1]
Hossein Abbasimehr, Mostafa Setak, Mohammad Tarokh	A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction [2]
М.Корыстов	Применение методов машинного обучения для предсказания поведения абонентов сотовой связи. [3]
А. Сулягина	Оптимизация предсказания оттока абонентов оператора сотовой связи [4]

В работе [1] рассматривается применение метода SVM для решения аналогичной задачи, изучается вопрос оптимизации за счет подбора параметров, а также проводится сравнение данного метода с другими алгоритмами машинного обучения: логистической регрессией и случайным лесом, при этом SVM показал средний результат.

В работе [2] рассматриваются способы построения ансамблей из различных базовых алгоритмов. В качестве последних были выбраны нейронные сети, решающие деревья и метод опорных векторов, из которых строились ансамбли с помощью бустинга, бэггинга, стекинга и простого голосования. В результате применения данных методов ансамблирования были получены более эффективные классификаторы, чем те, что были основаны исключительно на базовых алгоритмах. Лучший результат показал бустинг над решающими деревьями, однако показатели остальных алгоритмов оказались не сильно хуже. У SVM лучший результат был получен с применением бэггинга, поэтому на этот способ ансамблирования стоит обратить внимание.

В рассмотренных выше работах используются различные датасеты (наборы данных) с различными параметрами. Поэтому нельзя сказать, что задачи предсказания оттока, решаемые в [1] и [2] в точности те же, что рассматривается в данной работе. Однако, основываясь на их результатах, можно предположить, что с помощью метода опорных векторов можно добиться хорошей классификации и для данного датасета. В [3] и [4] метод опорных векторов не рассматривается, однако, эксперименты, описанные в данных работах проводились на схожих данных. Поэтому их лучший результат будет использоваться для сравнений. В этих работах были рассмотрены возможности логистической регрессии, нейронных сетей, бустинга над решающими деревьями, случайного леса и метода ближайших соседей. Лучший результат показал бустинг (precision = 0.75, recall = 0.72 AUC = 0.92).

5. Эксперимент

5.1. Данные

Для обучения алгоритмов использовались данные 300000 клиентов, предоставленные одним из крупнейших операторов сотовой связи в России. Данные содержали информацию об активности пользователей в течение четырех месяцев, а также информацию о том, ушел абонент или нет. Данные были агрегированы по следующим категориям:

- Количество минут и стоимость исходящих;
- Количество минут и стоимость входящих;
- Количество и стоимость исходящих СМС;
- Количество и стоимость входящих СМС;
- Количество трафика мобильного интернета и его стоимость;
- Информация о количестве обращений клиента в справочные службы;
- Идентификатор используемого тарифного плана;
- Личные данные.

5.2. Кросс-валидация

Для получения достоверной оценки эффективности классификатора использовался метод кросс-валидации [6]. Весь датасет был перемешан и разбит на десять непересекающихся частей. Затем поочередно каждая из частей выступает в качестве тестовой выборки, в то время как остальные девять используются для обучения алгоритма. На тестовой выборке с помощью введенных метрик оценивается качество классификации. Окончательный результат является усредненным значением качества всех десяти итераций.

5.3. Реализация

Вся работа была выполнена на языке Python. Были использованы следующие библиотеки:

- *Pandas* - для обработки данных;
- *Scikit-Learn* - для построения алгоритмов и оценки их эффективности;
- *Multiprocessing* - для распараллеливания алгоритма бэггинга;
- *Numpy* - для проведения трудоемких вычислений.

6. SVM

6.1. Описание метода

SVM (метод опорных векторов) - метод, применяемый в задачах бинарной классификации, заключающийся в поиске разделяющей гиперплоскости между двумя классами (C1, C2, Рис. 2). Главной особенностью данного метода является то, что в случае линейной разделимости выборки, он ищет гиперплоскость (H, Рис. 2) с максимальной шириной разделяющей полосы (Margin, Рис. 2), то есть разделяющую гиперплоскость [5], максимально отдаленную от обоих классов. В случае линейной неразделимости вводятся дополнительные переменные, характеризующие допустимую ошибку классификации на различных объектах. Также для линейно неразделимой выборки применяется трюк, заключающийся в переходе от скалярного произведения к нелинейной функции ядра (kernel trick). В качестве ядра может выступать любая функция, представимая в виде скалярного произведения в некотором пространстве. Данный прием позволяет перейти в пространство большей размерности, где выборка может быть линейно разделима.

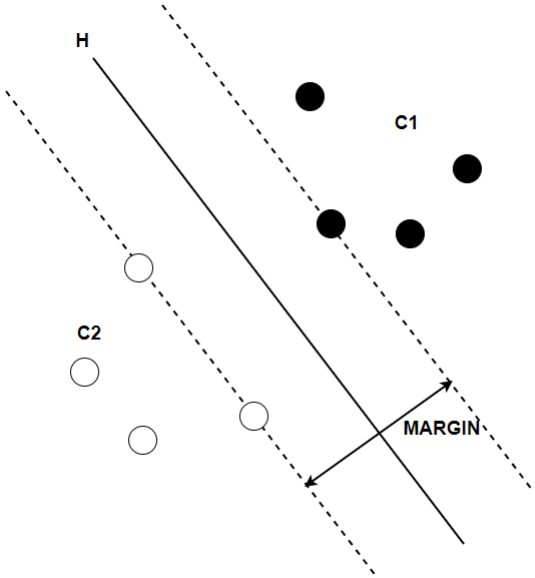

Рис. 2

6.2. Результаты

Как было показано в работах [1] и [2], в тех случаях, когда сложно определить, какое ядро даст лучшую точность, наиболее оптимальным выбором будет радиальное ядро (RBF [5]). Поэтому для экспериментов было выбрано именно оно.

$$K(x, x') = exp\{-\gamma||x - x'||^2\}$$
RBF ядро, $\gamma-$ внешний параметр

Результаты, полученные в ходе применения метода опорных векторов, приведены в таблице ниже. На рисунке (Рис. 3) отображена полученная ROC - кривая.

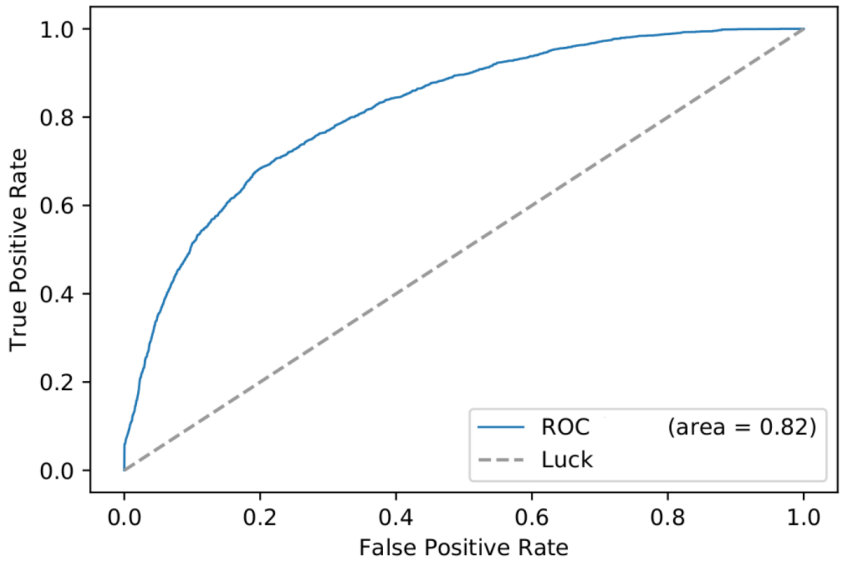

Рис. 3

AUC	Precision	Recall	F_1
0.82	0.58	0.43	0.49

Как видно, полученный результат уступает результату, выбранному для сравнения. Поскольку попытки улучшения качества предсказания за счет изменения параметров не дали положительного эффекта, было принято решение построить ансамбль из SVM.

7. Bagging

7.1. Описание метода

Идея данного метода заключается в том, чтобы независимо обучить несколько базовых алгоритмов на различных подвыборках данных, которые затем объединяются в композицию с помощью голосования. При этом подвыборки могут пересекаться, а также некоторые образцы из исходного датасета могут не попасть ни в одну подвыборку. В качестве метода определения окончательного решения обычно применяются:

- Простое голосование (Simple Voting) - конечный результат определяется как среднее среди ответов базовых алгоритмов;
- Взвешенное голосование (Weighted Voting) - конечный результат определяется как сумма ответов базовых алгоритмов с некоторыми ненулевыми весами (сумма весов должна быть равна единице).

Стоит также отметить, что в связи с независимостью базовых алгоритмов, их обучение может проводиться параллельно.

7.2. Результаты

Для определения результата классификации использовалось взвешенное голосование. То есть, если $\{b_k\}_{k=1}^N$ - множество независимо обученных алгоритмов, ω_k - весовые коэффициенты, то результирующий алгоритм получается следующим образом:

$$b(x) = \sum_{k=1}^{N} \omega_k b_k(x), \quad \sum_{k=1}^{N} \omega_k = 1, \omega_k \geq 0$$

Как и в случае с одним SVM, для каждого отдельного классификатора использовалось радиальное ядро. Метод опорных векторов с данным ядром имеет два внешних параметра: C и γ. Для выбора оптимальной конфигурации, была вычислена оценка AUC для различных комбинаций C и γ. Результаты приведены в таблице ниже.

$\gamma \setminus C$	1e-3	1e-1	1	1e3	1e5
1e-3	0.92	0.93	0.93	0.93	0.91
1e-1	0.95	0.96	0.95	0.95	0.94
1	0.89	0.91	0.91	0.93	0.92
1e3	0.94	0.93	0.94	0.92	0.94
1e5	0.93	0.93	0.95	0.92	0.92

Как видно из таблицы, при $\gamma = 0.1$ результат близок к наилучшему практически при любом значении C. Поэтому в качестве оптимальных параметров была выбрана пара $(0.1, 0.1)$. Полученные результаты представлены на изображении (ROC - кривая, Рис. 4) и в таблице ниже.

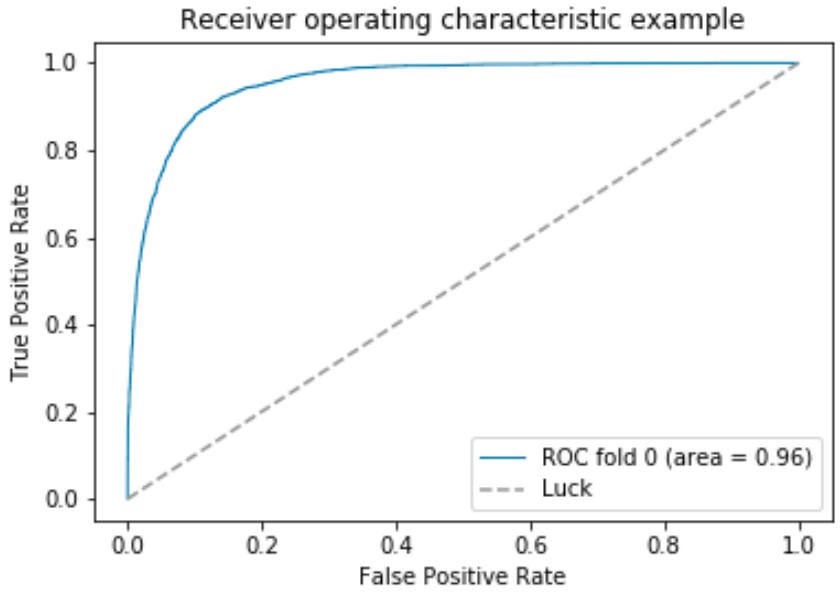

Рис. 4

AUC	Precision	Recall	F_1
0.96	0.76	0.77	0.76

Полученный результат превосходит сравниваемый. Поэтому можно заключить, что ансамбль из алгоритмов, основанных на методе опорных векторов применим для решения поставленной задачи.

Заключение

По предоставленной выборке были построены классификаторы на основе метода опорных векторов, а также на основе ансамблей SVM, построенных методом бэггинга. С помощью ансамблирования получилось достичь лучших результатов, чем в сравниваемых работах, поэтому данный метод можно считать применимым в задаче предсказания оттока абонентов.

Список литературы

[1] Kristof Coussement, Dirk Van den Poel. *Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques*, Expert Systems with Applications 34, 313–327, 2008.

[2] Hossein Abbasimehr, Mostafa Setak, and Mohammad Tarokh. *A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction*, The International Arab Journal of Information Technology, Vol.11, No.6, 2014.

[3] М. А. Корыстов. *Применение методов машинного обучения для предсказания поведения абонентов сотовой связи*, дипломная работа, 2015.

[4] А. А. Сулягина. *Оптимизация предсказания оттока абонентов оператора сотовой связи*, курсовая работа, 2015.

[5] К. В. Воронцов. *Лекции по методу опорных векторов*, 2007.

[6] Ю. С. Шунина. *Влияние способа формирования обучающей и тестовой выборок на качество классификации*, Вестник Ульяновского государственного технического университета, no. 2 (70), 2015, pp. 43-46.