Интеллектуальный анализ данных для задач CRM

Симонова С.И.

***Аннотация*—В данной статье рассмотрены методы интеллектуального анализа данных и возможность их применения организациями, использующими системы класса CRM, на примере проблемы повышения продаж. Проблема сведена к задаче бинарной классификации и решается при помощи современных алгоритмов классификации (дерево решений, градиентный бустинг, логистическая регрессия, нейронные сети). В работе приводятся описания и оценка качества построенных моделей классификаторов.**

***Ключевые слова*—интеллектуальный анализ данных, машинное обучение, классификация, маркетинг.**

I. Введение

На сегодняшний день почти все организации потребительской сферы (особенно в сфере розничных продаж) функционируют в условиях жесткой конкуренции, что требует грамотной работы с потребителями. Работа подразумевает деятельность в двух направлениях: привлечение новых клиентов и эффективное взаимодействие с настоящими. Задачи не являются взаимоисключающими, и сложно выделить какую-либо более важной, поэтому они решаются параллельно, с использованием различных методов.

Однако, кампании, направленные на поддержание взаимоотношений с клиентами, требуют затрат сравнимо меньших, чем кампании по привлечению клиентов. В то же время, клиент, с которым уже взаимодействовали ранее, приносит большую прибыль, чем клиент, новый для компании. Логическим выводом из обозначившейся закономерности явилась политика ориентации на клиента. Организации должны знать своих клиентов, и в настоящее время это стало аксиомой на рынке товаров и услуг с высокой конкуренцией.

В распоряжении современных компаний находится множество техник и методов эффективного взаимодействия с клиентами. Одна из таких техник, up-sale («повышение продаж»), состоит в наделении продаваемого продукта или услуги теми или иными добавочными характеристиками, которые придают им дополнительную ценность и склоняют клиента к покупке именно такого усовершенствованного продукта или услуги.

Статья получена 18 декабря 2014. Работа представляет собой результат магистерской диссертации. Симонова Сабина Игоревна, ich.bin.sabin@gmail.com, факультет ВМК МГУ имени М.В.Ломоносова

Очевидно, что с ростом клиентской базы, увеличением потока информации и объемов хранимых данных становится невозможным функционирование без информационных систем, предназначенных для автоматизации соответствующих процессов. В связи с этим широкое распространение получили системы класса CRM (Customer Relationship Management), и с каждым годом количество компаний, внедряющих системы CRM, растет.

Для поддержки этих бизнес-целей CRM-система может включать в себя несколько частей: фронтальная часть, обеспечивающая обслуживание клиентов на точках продаж, операционная часть, предназначенная для авторизации операций и оперативной отчетности, хранилище данных и аналитическая подсистема. Составная структура системы на каждом этапе обеспечивает слаженную работу всех элементов процесса.

В аналитических системах для решения маркетинговых задач используются достижения дисциплин статистики, машинного обучения, искусственного интеллекта, визуализации, теории информации, баз данных, объединенных термином интеллектуальный анализ данных (data mining).

Маркетинговые задачи сводятся к задачам интеллектуального анализа данных, таким как

1. Регрессия – нахождение функциональной зависимости между входными параметрами и непрерывным выходным параметром. Позволяет оценивать вероятность события или его численное значение. (Решается в задачах прогнозирования спроса, оценки ценовой эластичности, оценка вероятности повторных продаж, расчета загруженности склада, магазина, кассы, анализа влияния различных факторов на спрос)

2. Классификация – нахождение функциональной зависимости между входными параметрами и дискретным выходным параметром. Классификация позволяет отнести объект к одному из известных классов (Оценка перспективности клиентов; анализ рисков: стоит ли давать кредит или нет; оценка скидок: какой категории клиентов предоставлять скидки; прогнозирование успеха сделки, оценка эффективности рекламной компании).

3. Кластеризация – разбиение объектов на группы схожих элементов (кластеры). Позволяет анализировать одни объекты по аналогии с поведением других (Маркетинговые задачи: кластеризация товаров, выявление товаров со схожей структурой спроса,

разбиение клиентов на близкие по структуре и особенностям поведения группы, анализ спроса в зависимости от комбинации входных показателей, обнаружение аномальных отклонений).

4. Ассоциация – анализ транзакций, т.е. событий, происходящих вместе. Обнаружение зависимости, что из одного события с определенной вероятностью следует другое событие (Задача предсказание поведения клиента и предложение товара, который скорее всего его заинтересует, размещение товаров на полках, в каталогах, кросс-продажи – стимулирование продаж одних товаров за счет продажи других, оптимизация складских запасов).

II. Начальные сведения и определения

A. *Описание исходных данных*

Компания Orange Telecom предоставила базу клиентских данных. Набор данных содержит 50 000 наблюдений и 230 переменных, среди которых 190 числовых и 40 категориальных. Переменные характеризуются значительным количеством пропущенных значений, 31 переменная тренировочного набора является пустой, то есть в каждом из наблюдений не имеет значения.

В некоторых наблюдениях присутствуют пропущенные значения, что необходимо учитывать при подготовке данных для моделирования.

В целях защиты данных клиентов все названия переменных и их значения были анонимизированы. Таким образом, без дополнительной семантической информации невозможно сделать интуитивные или эвристические предположения при построении модели.

Целевая переменная набора данных распределена асимметрично в пользу отрицательных значений (в тренировочном наборе всего 7,4% записей имеют принадлежность к положительному классу), что затрудняет построение классификатора: чтобы обучить модель, в равной степени необходимы примеры записей, относящихся как к положительному, так и к отрицательному классу.

B. *Критерий оценки качества моделей*

Для оценки качества модели организаторы использовали ROC-кривую (receiver operating characteristic, операционная характеристика приёмника), наиболее часто используемую для представления результатов бинарной классификации. Поскольку различных классов всего два, один из них называют классом с положительными исходами, второй – с отрицательными исходами.

По оси абсцисс откладывают значения чувствительности алгоритма, sensitivity, по оси ординат – (specificity) значения специфичности алгоритма вычисляемые следующим образом:

$$\text{sensitivity} = \frac{tp}{tp+fn}, \qquad \text{specificity} = \frac{tn}{tn+fp},$$

где tp – число элементов, верно отнесенных к положительному классу, tn – число элементов, верно отнесенных к отрицательному классу, а fp и fn – число элементов, ложно отнесенных к положительному и отрицательному классу, соответственно [1].

Для сравнения моделей конкурсантов организаторы использовали численный показатель AUC (Area Under Curve), равный площади под ROC-кривой.

На практике, площадь под ROC-кривой принимает значения между 0.5 (тогда ROC совпадает с базовой линией) и единицей. В таблице 1 приведена интерпретация значений AUC[2].

Таблица 1. Интерпретация значений AUC

Значение AUC	**Характеристика модели**
0.5 – 0.6	слабая
0.6 – 0.7	удовлетворительная
0.7 – 0.8	хорошая
0.8 – 0.9	очень хорошая
0.9 – 1.0	превосходная

C. *Постановка задачи*

Задача работы состоит в построении модели прогнозирования повышения продаж При решении задачи будут использованы данные KDD Cup 2009, предоставленные компанией Orange Telecom, с сокращенным набором переменных (190 численных и 40 категориальных) и критерии оценки качества моделей аналогичные применявшимся в соревнованиях KDD Cup.

В качестве результата ожидается модель прогнозирования, имеющая значение AUC в интервале от 0.8 до 1.0 (характеристика модели «очень хорошая» или «превосходная»).

III. Исследование и построение залачи

Задача построения модели для задачи бинарной классификации состоит из нескольких частей:

1. Исследовать характеристики набора данных на предмет наличия корреляций, пропущенных значений, выбросов, изучить распределение переменных, в том числе целевой;
2. Провести предобработку данных, если требуется,: удаление ненужных прецедентов, подстановка пропущенных значений, фильтрация, удаление корреляций, группировка категориальных переменных, дискретизация непрерывных переменных;
3. Построить модель бинарной классификации;
4. Оценить качество модели и скорректировать в случае необходимости.

A. *Предобработка данных*

Данные, используемые для бизнес-анализа, чаще всего плохого качества. В них содержится много ошибок: дублирование, противоречия, пропуски, аномалии и множество других проблем. Исключить их полностью невозможно, поэтому данные нужно очищать.

Пропущенные значения могут повлиять на результаты анализа. Если игнорировать наличие пропусков в данных или полагать, что достаточно исключить из

анализа данные с пропущенными значениями, то существенно вырастает риск получения неверных или незначимых результатов.

Наиболее простым, и в тоже время, часто применимым методом является заполнение средним значением (модой, медианой или средним арифметическим - в зависимости от шкалы, по которой измерена переменная с пропусками), найденным по имеющимся данным. Средние значения, вычисленные на исходном и преобразованном массивах, совпадают. Однако такого рода преобразование «усредняет» данные, уменьшая дисперсию признака, а, следовательно, и показатели корреляции.

Существуют так же и более сложные методы, в своей основе содержащие регрессию, нейронные сети, метод ближайшего соседа [3].

Наборы накопленных данных могут содержать ошибки в виде дубликатов, шумов или выбросов. Наличие дубликатов в наборе данных может являться способом повышения значимости некоторых записей. Такая необходимость иногда возникает для особого выделения определенных записей из набора данных. Однако в большинстве случаев, продублированные данные являются результатом ошибок при подготовке данных (один и тот же объект внесен в справочник под различными названиями). Существует два варианта обработки дубликатов. В первом варианте удаляется вся группа записей, содержащая дубликаты. Этот способ используется в том случае, если наличие дубликатов вызывает недоверие к информации, полностью ее обесценивает. Второй вариант состоит в замене группы дубликатов на одну уникальную запись или объединении сведений по одинаковым объектам, с учетом того, что в значении могут быть опечатки, переставленные слова и прочие проблемы, не позволяющие проводить дедубликацию на основе полного совпадения [4].

В предоставленном наборе данных 230 переменных и 50000 записей. Такой объем обычно является избыточным, и наша задача состоит в том, чтобы выбрать те характеристики, которые наиболее полезны для классификации, и понизить размерность (уменьшить анализируемое множество данных до размера, оптимального с точки зрения решаемой задачи и используемой аналитической модели).

При подготовке наборов данных для интеллектуального анализа нередко применяют дискретизацию – процесс распределения значений непрерывного набора данных по сегментам так, чтобы получилось ограниченное число допустимых значений. Полученные сегменты воспринимаются как упорядоченные дискретные значения. Дискретизировать можно только числовые данные.

Применительно к номинальным переменным может использоваться процедура квантования (группировки) похожих значений для сокращения размерности данных, а именно для уменьшения числа разнообразных значений признака.

B. Построение модели бинарной классификации

Для решения задачи бинарной классификации существует ряд методов интеллектуального анализа данных. Далее рассмотрим наиболее применяемые.

Нейронные сети. представляют собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейронов). Такие процессоры обычно довольно просты, особенно в сравнении с процессорами, используемыми в персональных компьютерах. Каждый процессор подобной сети имеет дело только с сигналами, которые он периодически получает, и сигналами, которые он периодически посылает другим процессорам.

Нейронные сети не программируются в привычном смысле этого слова, они обучаются. Возможность обучения — одно из главных их преимуществ перед традиционными алгоритмами. Технически обучение заключается в нахождении коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными данными и выходными, а также выполнять обобщение. Это значит, что в случае успешного обучения сеть сможет вернуть верный результат на основании данных, которые отсутствовали в обучающей выборке, а также неполных и/или «зашумленных», частично искаженных данных [5].

Для решения задачи бинарной классификации широко применяют также и модели регрессии. В линейной регрессии зависимая и независимая переменная являются непрерывными (continuous). Метод находит линию, проходящую через данные и минимизирующую квадрат ошибок в каждой точке. Другими словами, это попытка найти наилучшую линейную функцию для заданных значений переменных. Логистическая регрессия во многом схожа с линейной. Ключевое отличие заключается в том, что зависимая переменная дискретная или категориальная, но не непрерывная. Это делает модель наиболее удобной для применения в маркетинге, потому как аналитики часто пытаются предсказать именно дискретное действие клиента, например, отклик на предложение или невыполнение обязательств по кредиту [6].

Еще один метод – дерево принятия решений. Дерево состоит из «листьев» и «веток». На ребрах дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе [6].

Наивный байесовский классификатор представляет собой простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими (наивными) предположениями о независимости.

В зависимости от точной природы вероятностной модели, наивные байесовские классификаторы могут обучаться очень эффективно. Во многих практических приложениях, для оценки параметров для наивных байесовых моделей используют метод максимального правдоподобия; другими словами, можно работать с наивной байесовской моделью, не веря в байесовскую вероятность и не используя байесовские методы.

Вопреки наивному виду и упрощенным условиям, наивные байесовские классификаторы часто работают намного лучше во многих сложных задачах. Достоинством наивного байесовского классификатора является малое количество данных для обучения, необходимых для оценки параметров, требуемых для классификации [7].

IV. Построение модели классификатора

В рамках поставленной задачи были построено пять моделей бинарной классификации с использованием деревьев решений, градиентного бустинга решающих деревьев, регрессии, нейронных сетей. Для решения задачи использовался инструмент SAS Enterprise Miner 13.2 (OnDemand).

В соответствии с методологией SEMMA (Sample, Explore, Modify, Model, Assess) первым этапом построения каждой из моделей стала выборка данных. При помощи узла File Import была произведена загрузка набора данных, узел Decisions используется для балансировки классов, узел Data Partition разделил входные данные на три набора с сохранением распределения – тренировочный, валидационный и тестовый – в соотношении 30:30:40. Таким образом, на 30% наблюдений классификатор будет обучены, следующие 30% для проверки качества модели и возможного изменения параметров, донастройки модели, а оставшиеся 40% – тестовые, используются для оценки модели.

Далее модели классификации имеют разные шаги, что обусловлено свойствами алгоритмов, их чувствительности к тем или иным недостаткам набора данных. Например, для классификатора деревья решений факт наличия пропущенных значений и большое количество атрибутов не является критичным, и можно построить модель, не прибегая к использованию узлов этапа Explore и Modify.

Согласно установленным параметрам классификатора, размер листа должен быть не меньше 5, максимальное количество правил узла не больше 5, пропущенные значения используются в классификации, критерий выбора параметра разделения для интервальных переменных ProbF, ProbChisq для номинальных и Entropy для порядковых; число ветвей 2, поскольку имеем дело с бинарной классификацией, максимальная глубина дерева не превышает 15 узлов.

Значимыми были приняты семь атрибутов: Var126, Var28, Var153, Var211, Var135, Var216, Var119. Результатом модели стало дерево, показавшее значения AUC на наборах тренировочный, проверочный, тестовый соответственно0.73, 0.73 и 0.74.

В следующей модели к дереву решений были добавлены узлы Start Groups и End Groups. Узлы полезны, в случае, если данные сегментированы или сгруппированы определенным образом, но есть необходимость провести перегруппировку. Первый узел определяет начало действия группировки, второй – ее завершение. Добавление узлов группировки значений атрибутов существенно повысило качество модели, результаты работы усовершенствованного алгоритма на наборах данных следующие: тренировочный – 0.99, валидационный – 0.83, тестовый – 0.82.

В диаграммы других трех моделей были включены узлы из блоков Explore (узел Variable Selection) и Modify (узлы Impute и Interactive Groupig). При помощи узла Variable Selection из 230 параметров можно выбрать те, которые оказывают наибольшее влияние на значение целевой переменной. Так в нашем случае узел проверяет гипотезы, основываясь на критерии хи-квадрат, причем нижняя граница значения 3.84, допускает не более 60% пропущенных значений у одной переменной, не более 100 классов категориальной переменной. По результатам работы узла были выбраны атрибуты Var126, Var28, Var211, Var204, Var13, Var38, Var207, Var212, Var189, Var206, Var132, Var44 в порядке убывания значимости. Более 60% пропущенных значений оказалось у 156 атрибутов, а в 11 атрибутах было превышено максимальное число значений классов. Остальные переменные не преодолели порог критерия хи-квадрат.

Таким образом, на вход узлу Impute были поданы наборы данных с числом атрибутов равным 12. Согласно установленным параметрам, допускается не более 60% пропусков в столбце атрибута, в пропущенные значения категориальных переменных будет подставлено наиболее часто встречающееся значение класса, в интервальные атрибуты будут подставлены средние значения вместо пропусков.

Следующий шаг диаграммы – узел Interactive Grouping, при помощи которого при необходимости значения переменных можно сгруппировать или, наоборот, разделить. Получившиеся переменные, ровно как и оставшиеся без изменения, можно подать на вход классификатору. Параметры узла установлены таким образом, что группировка происходит при помощи метода Bucket и с использованием 50 корзин.

Третьим классификатором стал градиентный бустинг деревьев решений, который согласно установленным параметрам, строится при помощи 10 деревьев и 10 итераций, предсказание каждого дерева уменьшается на 0.2, не более трех поколений узлов, минимальное допустимое число различных значений для категориальной переменной равно 5. Площадь под кривой ROC на трех наборах данных приняла значения 0.78, 0.79 и 0.79.

Четвертый построенный классификатор в своей основе имеет нейронную сеть с тремя слоями нейронов. Значения AUC для наборов train, validate, test равны соответственно 0.87, 0.86, 0.86.

Пятая модель классифицирует наблюдения из наборов при помощи узла Regression, согласно настройкам, использующего логистическую регрессию. Результатом работы модели стала классификация наблюдений с результатами, схожими модели с нейронной сетью: 0.87, 0.86 и 0.85.

Узел Model Compression, выбирающий из построенных моделей имеющую наибольшее значение AUC на тестовом наборе данных, определил, что модели с использованием нейронных сетей (модель №4) и

логистической регрессии (модель №5) показывают наилучшие результаты, незначительно отличающиеся друг от друга в пользу нейронной сети.

Модели, основанные на деревьях решений заметно уступают лидерам, но наиболее удачной, с качественно лучшим значением AUC (более 0.8) оказалась модель № 2, имеющая узлы группировки.

V. Заключение

Подведем итоги:

- рассмотрены способы применения интеллектуального анализа данных для решения маркетинговых задач, в частности, задачи повышения продаж;
- освещены наиболее эффективные и широко используемые методы решения задачи бинарной классификации;
- в рамках поставленной задачи на наборе данных соревнования KDD Cup 2009 построено пять моделей, три из которых имеют значение AUC в интервале от 0.8 до 0.9, а значит, характеризуют модель как «очень хорошая».

Библиография

[1] Gideon Dror, Marc Boulle, Isabelle Guyon, Vincent Lemaire, David Vogel *The 2009 Knowledge Discovery in Data Competition (KDD Cup 2009) Challenges in Machine Learning*, Volume 3.

[2] Mohamed Bekkar, Dr.Hassiba Kheliouane Djemaa, Dr.Taklit Akrouf Alitouche , *Evaluation Measures for Models Assessment over Imbalanced Data Sets*, Journal of Information Engineering and Applications, Vol.3, No.10, 2013, 27-38

[3] Hippel Paul T. *Biases in SPSS 12.0 Missing Value Analysis,* The American Statistician. May 2004. Vol. 58. No. 2.

[4] Чубукова И.А. *Data Mining*, 2006

[5] П.Г. Круг. *Нейронные сети и нейрокомпьютеры*, 2002

[6] Parr-Rud O. *Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management*, 2001

[7] Rish, Irina, *An empirical study of the naive Bayes classifier*, Proceedings of IJCAI-2001 workshop on Empirical Methods in AI (also, IBM Technical Report RC22230), pp. 41-46.

Data Mining for CRM Tasks

Simonova S.I.

Abstract— **The paper discusses data mining techniques and possibility to use them in CRM by companies. Marketing up-sell problem was reduced to the classification task and solved with actual classification methods (decision tree, gradient boosting, logistic regression, neural network). The article presents describing models and evaluation of the modeling results.**

Keywords—**Data mining, machine learning, classification, marketing.**