

УДК 519.2+61

ПРОГНОЗИРОВАНИЕ ИСХОДА ЭНДОВАСКУЛЯРНЫХ ОПЕРАЦИЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

М. С. Абрамович¹⁾, В. И. Стельмашок²⁾, Г. С. Дорофеев³⁾

^{1),3)} *Белорусский государственный университет, пр. Независимости, 4,
2203030, г. Минск, Беларусь, abramovichms@bsu.by*

²⁾ *Республиканский научно-практический центр «Кардиология»,
ул. Р. Люксембург, 110Б, 220036, г. Минск, Беларусь, stelval@yandex.by*

Обучающая выборка состоит из двух групп пациентов: тех, у кого операция прошла успешно, и тех, у кого возникло осложнение в виде перфорации коронарных артерий. С помощью точного критерия Фишера отбираются два набора информативных признаков. Для отбора признаков также применены метод LASSO и критерий взаимной информации. Предложен способ решения проблемы дисбаланса классов. Для решения задачи классификации используются алгоритмы машинного обучения, в том числе и ансамблевые методы. ROC-анализ показывает высокую точность классификации.

Ключевые слова: хронические тотальные окклюзии; перфорации коронарных артерий; машинное обучение; дисбаланс классов; бэггинг; бустинг.

PREDICTING THE OUTCOME OF ENDOVASCULAR OPERATIONS USING MACHINE LEARNING METHODS

M.S. Abramovich¹⁾, V.I. Stelmashok²⁾, G.S. Dorofeev¹⁾

^{1),3)} *Belarusian State University, Nezavisimosti Av., 4, 2203030, Minsk, Belarus,
abramovichms@bsu.by*

²⁾ *National Scientific and Practical Centre «Cardiology», R. Luksemburg 11B,
220036, Minsk, Belarus, stelval@yandex.by*

The training set consists of two groups of patients: those who had a successful surgery and those who ended up getting perforations of coronary arteries. Two sets of informative features are selected by using Fisher's exact test. The LASSO method and the mutual information criterion are also tried for feature selection. We also proposed the technique for handling class imbalance problem. Machine learning algorithms as well as ensemble methods are used for the classification problem solution. ROC analysis indicates high classification accuracy.

Keywords: chronic total occlusions; coronary artery perforations; machine learning; class imbalance; bagging; boosting.

Введение

Среди сердечно-сосудистых заболеваний лидирующие позиции занимает ишемическая болезнь сердца, возникающая вследствие несоответствия функционирования системы доставки (коронарных артерий) и реальной потребностью сердечной мышцы в кислороде.

Хронические тотальные окклюзии являются одним из наиболее часто встречаемых причин нарушения проходимости коронарных артерий у лиц, страдающих ишемической болезнью сердца. Согласно данным статистической отчетности, они верифицируются у 15-20% пациентов в ходе проведения коронарографии.

Установлено, что успешное восстановление кровотока в хронически окклюзированной коронарной артерии рентгенэндоваскулярными способами позволяет улучшить качество жизни. Вместе с тем проведение подобных вмешательств нередко ассоциируется с развитием осложнений, наиболее тяжелым из которых является перфорация стенки коронарного сосуда.

Разработка дооперационного прогнозирования риска развития перфораций в ходе реканализации хронических тотальных окклюзий коронарных артерий позволит предотвратить данное осложнение, оптимизировать качество оказания медицинской помощи вышеобозначенным пациентам.

Выборки состоят из классов успешно прооперированных пациентов и пациентов, прооперированных с перфорацией коронарных артерий. Необходимо отметить, что классы обучающих выборок имеют различный объем (несбалансированы).

В настоящее время решающие правила классификации строятся с использованием ансамблей алгоритмов классификации, основанных на следующих технологических подходах: бэггинг, бустинг и стекинг. Эти подходы позволяют повысить точность классификации по сравнению с отдельными классификаторами. При построении решающих правил определены значения гиперпараметров моделей, которые обеспечивают наибольшую точность классификации пациентов.

На основе обработанных данных ангиографических и клинических характеристик в дооперационный период будут определяться пациенты, которым с высокой эффективностью и безопасностью может быть выполнена реканализация хронических окклюзионных поражений коронарных артерий.

Описание выборки прооперированных на коронарных артериях пациентов

Хронические тотальные окклюзии коронарных артерий (ХТО) являются достаточно часто встречающимися поражениями коронарного русла и регистрируются у 20-30% лиц, имеющих верифицированную или предполагаемую ишемическую болезнь сердца [1,2]. Восстановление нормального антеградного кровотока в питающем сердечную мышцу хронически окклюзированном сосуде несет несомненную пользу, так как в данном случае улучшается как качество жизни, так и прогноз оперируемого пациента [3,4]. В то же время выполнение лечебных рентгенэндоваскулярных вмешательств представляется технически более сложным и несет потенциально более высокие риски развития осложнений по сравнению с аналогичными манипуляциями, проводимыми в неокклюзированных коронарных сосудах.

Перфорация коронарных артерий – одно из наиболее грозных осложнений, встречающихся в процессе реканализации ХТО коронарных артерий; в ряде случаев она ассоциируется с развитием тампонады, потребности в выполнении экстренного кардиохирургического вмешательства и смерти пациента [5].

Настоящее исследование являлось ретроспективным, одноцентровым и суммировало данные 397 пациентов, которым предпринималась попытка реканализации ХТО коронарных артерий антеградным доступом. В зависимости от наличия, либо отсутствия возникновения отмеченного осложнения в ходе антеградного проведения коронарного проводника через зону хронического окклюзионного поражения, все пациенты были разделены на 2 группы: лица у которых была зарегистрирована перфорация ($n = 28$) и пациенты, у которых не было отмечено данного осложнения ($n = 369$).

Отбор информативных признаков

Выбор полезной информации (отбор признаков) является важной операцией, поскольку для решения любой классификационной задачи целесообразно отобрать признаки, несущие не «шум» и не иррелевантную (не относящуюся к цели исследования), а полезную для данной задачи информацию. Отбор признаков сокращает стоимость сбора информации. В практических задачах затраты на измерение или вычисление отдельных признаков могут быть сопоставимы со стоимостью потерь от ошибочных прогнозов.

Отбор информативных признаков проводился с помощью метода LASSO (LeastAbsoluteShrinkageandSelectionOperator). Метод LASSO решает проблему переобучения из-за большого количества признаков [6]. Регрессия LASSO направлена на выявление переменных и соответствующих им коэффициентов регрессии, которые приводят к модели минимизирующей ошибку предсказания. Это достигается путем наложения ограничения на параметры модели, которые «сжимают» коэффициенты регрессии до нуля, то есть заставляют сумму абсолютных значений коэффициентов регрессии быть меньше фиксированного значения λ . В практическом смысле это ограничивает сложность модели. Переменные с нулевым коэффициентом регрессии после сжатия исключаются из модели. Модель LASSO может быть описана следующими формулами:

$$(\hat{\alpha}, \hat{\beta}) = \arg \min \left\{ \sum_{i=1}^N \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\},$$

$$\sum_j |\beta_j| \leq \lambda,$$

где x_{ij} – значение i -го наблюдения j -го признака; α и β – коэффициенты регрессии, y_i – метка класса, относящаяся к i -му наблюдению, λ – гиперпараметр, выбираемый заранее.

Выбор λ производится с помощью k -кратной перекрестной проверки. Для этого подхода набор данных случайным образом разбивается на k подвыборок равного размера. В то время как $k-1$ подвыборки используются для обучения модели прогнозирования, оставшаяся подвыборка используется для проверки этой модели. Эта процедура выполняется k раз, причем каждая из k подвыборок используется для проверки. Создается общий результат путем объединения k отдельных результатов проверки для диапазона значения λ и выбирается значение λ с лучшим результатом, которое затем используется для определения окончательной модели.

Мы также использовали точный критерий Фишера и критерий взаимной информации для отбора признаков, но в итоге модели, обученные на признаках, отобранных с помощью этих стратегий, показали худший результат.

Проблема несбалансированности классов

Проблема несбалансированности классов решается в настоящее время с использованием двух общих стратегий. Первая, которая является

независимой от классификатора, состоит из балансировки исходного набора данных с использованием различных подходов сэмплинга. Наиболее распространенными методами балансировки классов являются: уменьшение объема большего класса (undersampling) и увеличение объема меньшего класса (oversampling).

Вторая стратегия заключается в модификации классификаторов для их адаптации к наборам данных.

Восстановление баланса классов может проходить двумя путями. В первом случае удаляют некоторое количество наблюдений мажоритарного класса (undersampling), во втором – увеличивают количество наблюдений миноритарного класса (oversampling).

Самой простой стратегией является стратегия удаление примеров мажоритарного класса. Для этого рассчитывается число k – количество мажоритарных примеров, которое необходимо удалить для достижения требуемого уровня соотношения различных классов. Затем случайным образом выбираются k мажоритарных примеров и удаляются. Примеры из мажоритарного класса могут удаляться не только случайным образом, но и по определенным правилам.

Другой подход – это дублирование примеров миноритарного класса. В зависимости от того, какое соотношение классов необходимо, выбирается количество случайных наблюдений для дублирования.

Такой подход к восстановлению баланса не всегда может оказаться самым эффективным, поэтому был предложен специальный метод увеличения числа примеров миноритарного класса – алгоритм SMOTE (Synthetic Minority Oversampling Technique) [7]. SMOTE – это метод, который можно использовать для передискретизации данных. Эта техника создает новую синтетику вместо передискретизации путем замены. SMOTE вводит синтетические примеры в сегментах линии для передискретизации миноритарных отсчетов класса. Она объединяет все классы k -миноритариев, которые находятся близко к соседям. Выбор соседей k -ближайших соседей случаен. Число зависит от количества передискретизации, необходимого модели.

Классификация успешно прооперированных и прооперированных с перфорациями коронарных артерий пациентов

Построение решающих правил классификации успешно прооперированных и прооперированных с перфорациями коронарных артерий пациентов проводилось с использованием следующих методов машинного обучения: логистической регрессии, метода опорных векторов, случайного леса и градиентного бустинга [8,9].

Для обучения и тестирования моделей был использован метод скользящего экзамена (leaveoneout). Модель обучается на всех объектах, кроме одного, и выдает предсказание для оставшегося. Так происходит со всеми объектами и, таким образом, набирается предсказание для каждого объекта в выборке. Учитывая наличие истинных меток, можно вычислять различные метрики, такие как точность, полнота и F1-мера [8]. При этом каждый раз обучение происходит практически на всей выборке, а оценка эффективности модели на тесте более надежна, чем в случае одинарного разбиения на обучающую и тестовую выборки.

Для исключения дисбаланса классов применялась процедура undersampling в комбинации с процедурой oversampling. В качестве процедуры oversampling был применен алгоритм SMOTE, а в качестве процедуры undersampling – алгоритм RepeatedEditedNearNeighbors. Его суть заключается в следующем: для каждого объекта алгоритм k -ближайших соседей пытается определить метку класса. Если объект принадлежит мажоритарному классу, а алгоритм k -ближайших соседей присваивает метку миноритарного класса, то объект удаляется из выборки. Если объект принадлежит миноритарному классу и алгоритм k -ближайших соседей ошибается, то удаляются наблюдения мажоритарного класса, являющиеся соседями объекта.

Если используется процедура oversampling, то необходимо убедиться в том, что на кросс-валидации синтетические объекты ни разу не попадут в тестовую выборку. Иначе полученная оценка может быть завышенной. Таким образом, leaveoneout кросс-валидация проводится только на оригинальных объектах

Подбор гиперпараметров алгоритмов осуществлялся по сетке с использованием перекрестной проверки по отдельным объектам. Таким образом, была использована процедура вложенной кросс-валидации.

Во всех случаях проводились процедуры undersampling с помощью алгоритма RepeatedEditedNearestNeighbors из библиотеки imblearn (параметр k равен двум) и oversampling с помощью алгоритма SMOTE с параметрами по умолчанию.

Методом LASSO было отобрано 27 информативных признаков. Коэффициент регуляризации равен 0.7. Количество наблюдений в обоих классах (после применения алгоритмов SMOTE и EditedNearNeighbors) – 291.

В табл. 1 представлены оптимальные гиперпараметры для рассмотренных моделей, а в табл. 2 – точность, полнота, F1-мера и показатель AUC для набора информативных признаков, отобранных методом LASSO. Кросс-валидация по отдельным объектам проводилась только для оригинальных объектов (291 объектов мажоритарного класса и 28 – миноритарного).

Таблица 1

Гиперпараметры классификаторов для набора информативных признаков, отобранных методом LASSO

Метод	Оптимальные параметры
Случайный лес	Максимальная глубина дерева=7 Число деревьев=130
Логистическая регрессия	C=0.5
Метод опорных векторов	C=20 Гамма = 0.1
Градиентный бустинг	Скорость обучения =0.1 Максимальная глубина дерева =11 Число листьев =21

Таблица 2

Метрики точности классификации для набора информативных признаков, отобранных методом LASSO

Метод	Точность	Полнота	F1-мера	Показатель ROC-AUC
Случайный лес	0.57	0.82	0.68	0.94
Логистическая регрессия	0.40	0.68	0.51	0.89
Метод опорных векторов	0.64	0.64	0.64	0.96
Градиентный бустинг	0.68	0.82	0.74	0.97

Как следует из табл. 2, наибольшая точность классификации для набора информативных признаков, отобранных методом LASSO была достигнута градиентным бустингом.

Необходимо отметить, чтобы при применении процедуры *undersampling*, из выборки удалялось минимальное количество наблюдений. Если это требование удовлетворяется, а метрики классификации становятся лучше, то предполагается, что удаленные наблюдения были шумовыми, и их удаление будет способствовать лучшей классификации.

На рис. 1 представлены ROC-кривые классификаторов для набора информативных признаков, отобранных методом LASSO.

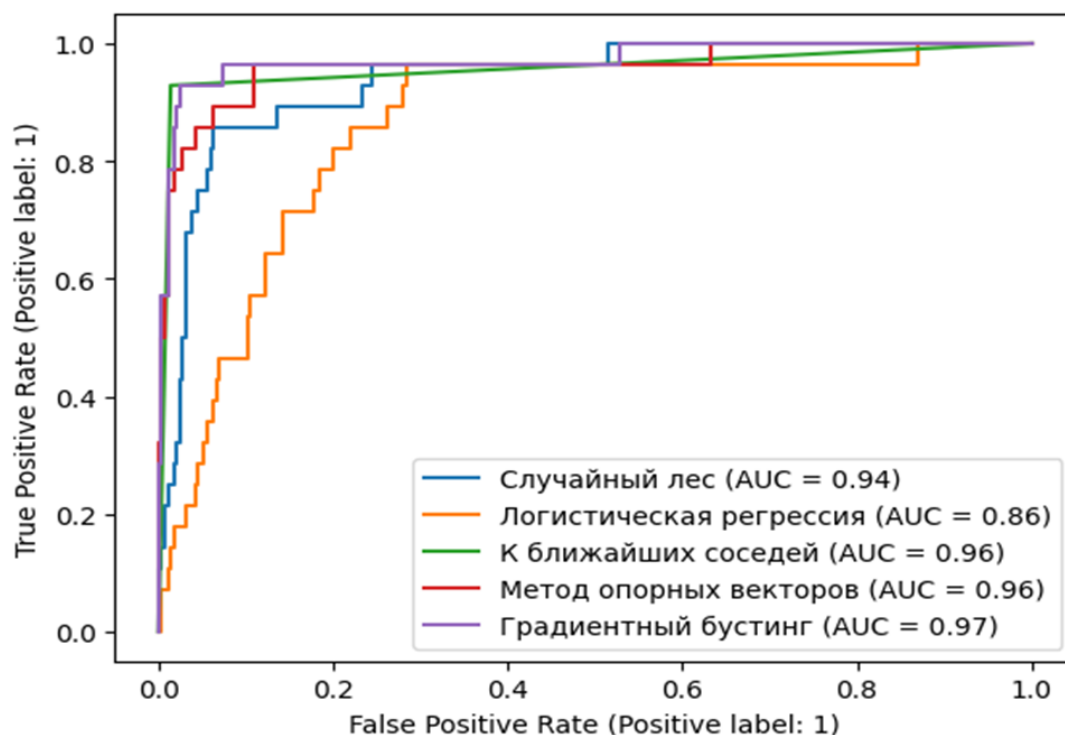


Рис. 1. ROC-кривые классификаторов для набора информативных признаков, отобранных методом LASSO

Оценка важности признаков классификации

После построения модели градиентного бустинга с оптимальными параметрами был проведен анализ важности признаков. Это позволяет провести интерпретацию модели.

Отдельные деревья решений можно интерпретировать, визуализируя их структуру. Но в модели градиентного бустинга содержатся сотни деревьев, и поэтому её нелегко интерпретировать путем визуализации входящих в неё деревьев. При этом необходимо понимать, какие именно признаки в данных оказывают наибольшее влияние на предсказание композиции.

Можно отметить следующее: признаки, используемые в верхней части дерева решений, влияют на окончательное предсказание для большей доли обучающих объектов, чем признаки, попавшие на более глубокие уровни. Таким образом, ожидаемая доля обучающих объектов, для которых происходило ветвление по данному признаку, может быть использована в качестве оценки его относительной важности для итогового предсказания. Усредняя полученные оценки важности признаков по всем решающим деревьям из ансамбля, можно уменьшить дисперсию такой оценки и использовать ее для отбора признаков.

Подход, который был нами применен, состоит в том, что важность отдается признаку, при использовании которого при разделении дерева решений происходит больший прирост точности.

Следует отметить, что оценка важности проводилась лишь для клинико-демографических признаков, измеренных в дооперационный период. Результаты представлены на рис. 2.

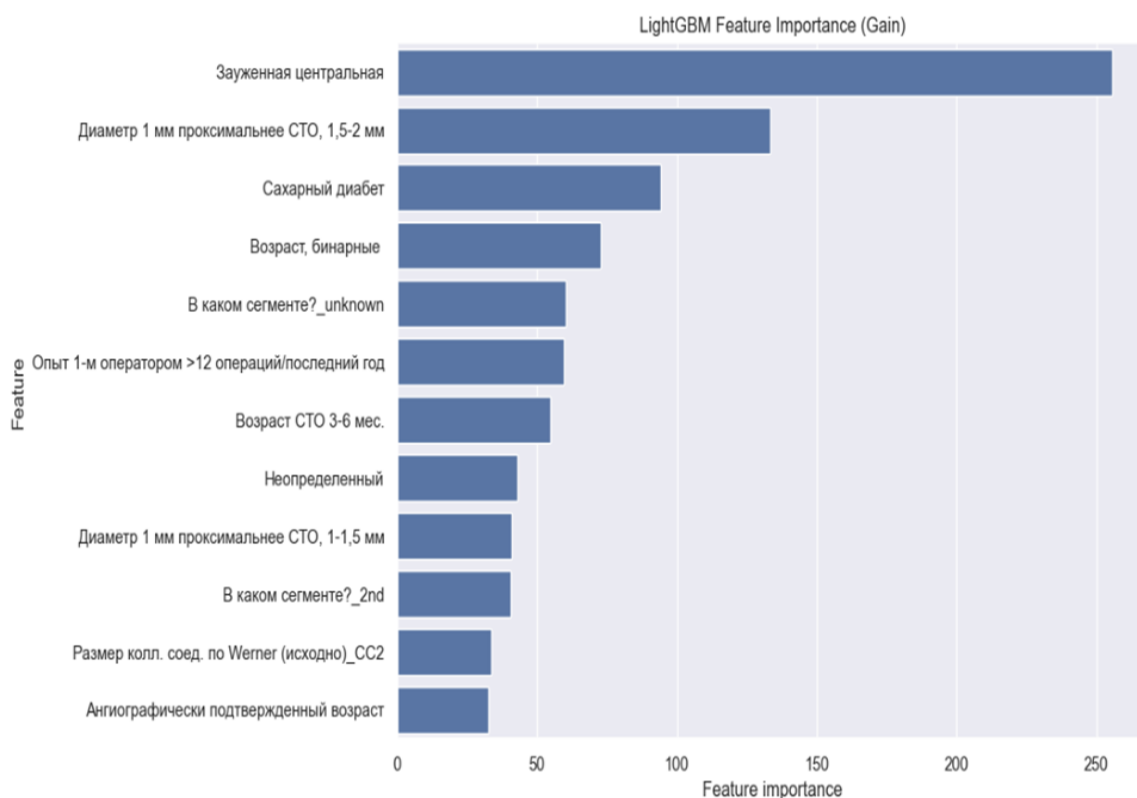


Рис. 2. Важность признаков для классификатора градиентный бустинг с отбором информативных признаков методом LASSO

Прогнозирование дооперационного риска перфораций коронарных артерий

Рассматривались только клинико-демографические признаки, измеренные в дооперационный период.

Для прогнозирования дооперационного риска перфорации коронарных артерий отбор информативных признаков проводился методом LASSO.

Оптимальное значение коэффициента регуляризации – 0.7. Параметр $n_neighbors$ в алгоритме RepeatedEditedNearNeighbors – 2. Количество наблюдений в обоих классах после проведения процедур undersampling и oversampling – 293.

В табл. 3 представлены найденные оптимальные гиперпараметры для рассмотренных моделей, а в табл. 4 – точность, полнота и F1-мера, а также показатель AUC для набора информативных признаков.

Таблица 3

Гиперпараметры классификаторов для набора дооперационных информативных признаков, отобранных методом LASSO

Метод	Оптимальные параметры
Случайный лес	Максимальная глубина дерева=11 Число деревьев=100
Логистическая регрессия	C=5
Метод опорных векторов	C=5 Гамма= 0.1
Градиентный бустинг	Скорость обучения= 0.5 Максимальная глубина дерева = 7 Число листьев = 21

Таблица 4

Метрики точности классификации для набора дооперационных информативных признаков, отобранных методом LASSO

Метод	Точность	Полнота	F1-мера	Показатель AUC
Случайный лес	0.60	0.64	0.62	0.96
Логистическая регрессия	0.36	0.57	0.44	0.84
Метод опорных векторов	0.88	0.50	0.64	0.98
Градиентный бустинг	0.82	0.75	0.78	0.98

Как следует из табл. 4, по F1-мере лучшей моделью для набора информативных признаков, отобранных методом LASSO, является градиентный бустинг.

Заключение

Таким образом, в задаче классификации пациентов нам удалось добиться высоких результатов даже при условии дисбаланса классов обучающей выборки и малого количества примеров миноритарного класса в целом. При отборе признаков лучшей оказалась стратегия, использующая модель L1-регуляризованной логистической регрессии. При устранении дисбаланса классов лучше всего себя показала комбинация алгоритмов SMOTE и EditedNearNeighbors. Лучшей моделью классификации оказался градиентный бустинг. Значения метрик F1 и ROC-AUC позволяют судить о том, что модель обладает высокой предсказательной способностью.

Библиографические ссылки

1. European perspective in the recanalisation of Chronic Total Occlusions (CTO): consensus document from the EuroCTO Club / C. Di Mario [et al.] // *EuroIntervention*. 2007. Vol. 3, № 1. P. 30–43.
2. In-hospital outcomes of percutaneous coronary intervention in patients with chronic total occlusion: insights from the ERCTO (European Registry of Chronic Total Occlusion) registry / A. Galassi [et al.] // *EuroIntervention*. 2011. Vol. 7, № 4. P. 472–479.
3. A randomized multicentre trial to compare revascularization with optimal medical therapy for the treatment of chronic total coronary occlusions / G. S. Werner [et al.] // *Eur. Heart J.* 2018. Vol. 39, № 26. P. 2484–2493.
4. Successful revascularization improves long-term clinical outcome in patients with chronic coronary total occlusion / T. Teramoto [et al.] // *Int. J. of Cardiol. Heart & Vasc.* 2017. Vol. 14. P. 28–32.
5. Angiographic success and procedural complications in patients undergoing percutaneous coronary chronic total occlusion interventions: a weighted metaanalysis of 18,061 patients from 65 studies / V. Patel [et al.] // *JACC Cardiovasc. Interv.* 2013. Vol. 6, № 2. P. 128–136.
6. *Ranstam J., Cook J.A.* LASSO regression // *Journal of British Surgery*. 2018. Vol. 105, № 10, P. 1348–1348.
7. *Chawla N.* SMOTE: Synthetic Minority Over-sampling Technique / N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer // *Journal of Artificial Intelligence Research*. 2002. Vol. 16. P. 341–378.
8. *Hastie T., Tibshirani R., Friedman J. H.* The Elements of Statistical Learning. 2nded. New York, Springer Publ. 2009.
9. *Harrington P.* Machine Learning in Action. New York: Manning. 2012.