Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

МЕТОД ОПОРНЫХ ВЕКТОРОВ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 441 группы

направления 09.03.03 Прикладная информатика

механико-математического факультета

Трусовой Натальи Алексеевны

Научный руководитель
д.ф.-м.н. _____ С.П.Сидоров

Заведующий кафедрой
д.ф.-м.н. _____ С.П.Сидоров

Саратов 2017г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современном мире любая сфера деятельности так или иначе связана с информацией и данными. Получение, хранение и обработка информации является неотъемлемой частью большинства процессов. В связи с этим происходит непрерывное накопление данных, что представляет определенную проблему обработки и анализа больших объемов информации. Все это влечет за собой активное развитие направлений автоматизации инструментов для работы с данными. Не меньшее развитие переживает область искусственного интеллекта и связанные с ней направления, например машинное обучение интеллектуальный анализ данных.

Машинное обучение представляет собой обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов для анализа данных и получения выводов и выноса решения или предсказания в отношении чего-либо. Подход, при котором прошлые данные или примеры используются для первоначального формирования и совершенствования схемы предсказания, называется методом машинного обучения. Общая задача машинного обучения заключается в восстановлении зависимости между входными и выходными элементами с целью предсказания будущего выхода по заданному входу. Целью машинного обучения является построение максимально точной модели на основе данных и затем использования этой модели для предсказаний в будущем.

Большая вариативность позволяет применять методы машинного обучения для данных различных типов в самых разных областях: биоинформатике, медицинской диагностике, технике, экономике. Так они используются для обнаружения мошенничества, кредитного скоринга, биржевого технического анализа. Одним из популярных направлений финансового анализа в последние годы является прогнозирование цен акций и поведения фондовых индексов на основе данных о предыдущих торговых периодах, т.к. положение на финансовых рынках оказывает непосредственное влияние на окружающую экономическую обстановку. Для получения релевантных результатов необходимы подходящие инструменты и корректные алгоритмы, в связи с чем методы машинного обучения и data mining получили широкое применение при анализе и прогнозировании финансовых рынков. Существует множество методов машинного обучения эффективно применяемых для данного

класса задач: искусственные нейронные сети, деревья принятия решений, логистическая регрессия, генетический алгоритм.

К таких методам принадлежит метод опорных векторов. Этот метод машинного обучения является набором схожих алгоритмов обучения с учителем, он также известен как машина опорных векторов (support vector machine, SVM) и используется для задач классификации и регрессионного анализа.

Востребованность данного метода заключается в том, что машина опорных векторов может обеспечить высокое качество обобщения, не обладая априорными знаниями о предметной области конкретной задачи [**?**], в связи с чем SVM-алгоритмы приобрели популярность и перспективные практические приложения.

Теоретический базис, уникальность идеи и интерпретируемость метода привлекли как исследователей, которые были заинтересованы в приложении этой теории в различных областях, таких как прогнозирование ситуации на финансовых рынках. Метод опорных векторов применяется для прогнозирования временных рядов, оценки кредитоспособности, а также в сфере финансовой безопасности.

В данной работе метод опорных векторов будет рассмотрен в рамках решения задачи классификации. Целью является изучение метода опорных векторов для случая бинарной классификации, и его применение для решения практической задачи - построить и тестировать простую стратегию управления активом, основанную на машинном обучении. Для достижения поставленной цели будут решаться следующие задачи:

1. Изучить теоретические основы метода опорных векторов.
2. Выбрать наиболее подходящие технологии реализации.
3. Реализовать модель обучения.
4. Применить реализованный алгоритм для анализа данных финансового рынка.
5. Проанализировать результаты обучения.

Первый раздел данной работы посвящен теоретической основе метода опорных векторов: математическому описанию метода и алгоритмам построения классификаторов.

В разделе 2 приведена реализация метода опорных векторов для прогнозирования движения цен акций и анализ полученных результатов.

1 Метод опорных векторов

Основы данного подхода были заложены в работах по статистической теории обучения в 1960-х годах. В своем первоначальном виде алгоритм решал задачу классификации объектов двух классов при линейно разделимой выборке, а в 1992 году был предложен способ адаптации машины опорных векторов для нелинейного разделения классов. К концу 20 века метод опорных векторов стал во много раз мощнее за счет использования ядерных функций, которые позволили значительно упростить задачу оптимизации и строить сложные разделяющие гиперплоскости, используя при этом лишь подвыборку данных - опорные вектора - для классификации тестовых объектов.

Задача классификации и регрессии с помощью метода опорных векторов, имеет целью разработку алгоритмически эффективных методов построения оптимальной разделяющей гиперплоскости в пространстве признаков высокой размерности.

Как было указано во введении, в рамках данной работы рассматривается применение метода опорных векторов для задачи классификации, когда выход принимает конечное число значений.

1.1 Оптимальная разделяющая гиперплоскость

Предварительно рассмотрим вариант полностью разделимой выборки, т.е. случай, когда обучение возможно провести без ошибок и разделить выборку на два непересекающихся класса. В контексте p-мерного пространства гиперплоскость представляет собой подпространство размерностью *p-1* и задается уравнением:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0. \quad (1)$$

Таким образом любая величина, в отношении которой справедливо уравнение (1), представляет собой точку этой гиперплоскости.

Допустим, что существует матрица X размерности n x p, содержащая n обучающих наблюдений в p-мерном пространстве, и что наблюдения относятся к двум классам, т.е. $y_1, \ldots, y_n \in \{-1, 1\}$. Тогда контрольное наблюдение представляет собой p-мерный вектор со значениями признаков $x^* = (x_1^*, ..., x_p^*$). В данном случае цель представляет создание классификатора на основе обучающих данных, который позволит правильно предсказать класс кон-

трольного наблюдения на основе его признаков.

Если разделяющая гиперплоскость существует, то возможно осуществить классификацию на основании того, с какой стороны от нее находится наблюдение. Если функция $f(x^*) = \beta_1 x_1^* + \beta_2 x_2^* + ... + \beta_p x_p^*$ положительна, то контрольное наблюдение будет отнесено к классу 1, если отрицательна – то к классу -1.

Оптимальная разделяющая гиперплоскость – единственная гиперплоскость, такая что сумма расстояний от ближайшей к ней точек выборки максимальна, среди всех разделяющих гиперплоскостей, расположенных на равных от них расстояниях.

Зазором или разделяющей полосой называется минимальное из расстояний от каждого наблюдения до некоторой разделяющей гиперплоскости. Способ классификации такого рода называется классификатором с максимальным зазором.

Обучающие наблюдения, которые находятся на одинаковом удалении от гиперплоскости с максимальным зазором и обозначают ширину зазора, являются опорными векторами. Т.е. опорные векторы представляют собой подмножество обучающей выборки, объекты которого лежат на границах разделяющей полосы. Очевидно, что гиперплоскость с максимальным зазором зависит только от опорных векторов и не зависит от остальных наблюдений. Сдвиг любого из них не повлиял бы на нее, если только в результате этого сдвига точка не пересечет границу зазора.

1.2 Классификация на основе опорных векторов

Описанная выше классификация возможна только при условии, что выборка наблюдений линейно разделима и разделяющая гиперплоскость существует, что не всегда возможно.

Чрезвычайно высокая чувствительность гиперплоскости с максимальным зазором предполагает большую вероятность переобучения – явления, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки будет существенно выше, чем средняя ошибка на обучающей выборке.

В данном случае необходимо рассмотреть классификацию на основе гиперплоскости, не выполняющей четкого разделения классов, что позволит обеспечить более высокую устойчивость к отдельным наблюдениям и более высокое качество классификации большинства обучающих наблюдений. Что-

бы обобщить алгоритм на случай линейной неразделимости, допустим возможность возникновения ошибки при классификации обучающих наблюдений.

В таком случае возможно построение разделяющей гиперплоскости с мягким зазором (soft margin) [?]. Зазор называется мягким, потому что некоторые точки могут нарушать его границы. Классификатор, основанный на данном подходе, называется классификатор на опорных векторах или классификатор с мягким зазором.

Вместо нахождения зазора максимальной ширины, при котором каждое наблюдение не только располагается с верной стороны от гиперплоскости, но и на правильной стороне относительно соответствующей границы зазора, допустим, что некоторые наблюдения могут находиться как на неправильной стороне относительно границы зазора, так в неверном классе – на неправильной стороне относительно гиперплоскости. На практике для построения SVM решают именно эту задачу, так как гарантировать линейную разделимость выборки в общем случае не представляется возможным.

Как сказано выше, классификатор на основе опорных векторов так же относит контрольное наблюдение к тому или иному классу в зависимости от того, по какую сторону относительно гиперплоскости оно находится. Гиперплоскость выбирается так, чтобы правильно разделять на два класса большинство обучающих наблюдений, но при этом допускается неверная классификация некоторой группы наблюдений. Это является решением следующей оптимизационной проблемы:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{максимизировать M}} \tag{2}$$

при условии, что

$$\sum_{j=1}^{p} \beta_j^2 = 1, \tag{3}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon), \tag{4}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C, \tag{5}$$

где M – это ширина зазора, которую нужно максимизировать, $\epsilon_1, ..., \epsilon_n$ - фиктивные переменные, которые позволяют отдельным наблюдениям находиться на неправильной стороне относительно границы зазора или относительно гиперплоскости, C – некоторый неотрицательный гиперпараметр, который задает допустимое число нарушений границы зазора или гиперплоскости и их выраженность.

Решающее правило классификатора на опорных векторах задается потенциально небольшим (в сравнении с объемом обучающей выборки) подмножеством наблюдений, следовательно, оно довольно устойчиво к поведению элементов выборки, расположенных далеко от гиперплоскости. Это делает его отличным, например, от решающего правила линейного дискриминантного анализа, которое, зависит от всех наблюдений в каждом классе, а так же от матрицы ковариаций, которая строится по всем имеющимся наблюдениям

1.3 Машины опорных векторов

Классификатор на опорных векторах представляет собой подход для распознавания объектов двух классов, когда решающая граница между классами является линейной. Так,наличие нелинейной связи между предикатом и откликом при использовании метода линейной регрессии может привести неудовлетворительным результатам. Для учете характера такой связи прибегают к расширению пространства, используя функции исходных предикторов, в частности квадратичные и кубические.

Машина опорных векторов (SVM) представляет собой обобщенную общую версию классификатора на опорных векторах, где реализована идея перехода в пространство большей размерности. Это достигается путем расширения пространства предикторов с помощью функций ядра. Использование рассматриваемых здесь ядерных функций является эффективным способом реализации этой идеи и будет рассмотрено ниже.

Для классификатора на основе опорных векторов решение проблемы (2) - (5) базируется на использовании только скалярных произведений наблюдений, а не самих исходных наблюдений. Таким образом решение для функции (**??**) можно представить в виде уравнения

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle, \tag{6}$$

.

При нахождении классификатора скалярное произведение заменим на следующую обобщенную форму скалярного произведения

$$K(x_i, x_{i'}) \tag{7}$$

где K - некоторая функция, называемая ядром.

Например, функция

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{8}$$

известна как линейное ядро, которое даст обычный классификатор на опорных векторах, поскольку он является линейным. Можно было бы выбрать другую форму для (7). Например

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d \tag{9}$$

- полиномиальное ядро степени d, где d - положительное целое число. Использование этого ядра при значениях d>1 вместо стандартного линейного ядра (8) в алгоритме классификатора на опорных векторах приводит к более гибкой решающей границе.

Широкое применение получило также радиальное ядро следующего вида:

$$K(x_i, x_{i'}) = exp(-\gamma + \sum_{j=1}^{p} x_{ij} x_{i'j})^2, \tag{10}$$

где γ - некоторая положительная константа. Рассмотрим некоторые особенности радиального ядра.

Отличие использования ядер от методов расширения пространства признаков при помощи функций исходных признаков заключается в том, что при использовании ядер требуется вычислить $K(x_i, x_{i'})$ только для всех $\binom{n}{2}$ уникальных пар (i, i') без непосредственной работы с расширенным пространством признаков. Данное обстоятельство является важным, поскольку во многих случаях применения SVM расширенное пространство признаков настолько велико, что вычисления становятся практически невозможными.

2 Классификации методом опорных векторов на языке R

Благодаря аналитическим возможностям метод опорных векторов вызывает постоянный интерес как инструмент для прогнозирования и классификации самых в различных сферах применения, в том числе и в финансах, при том, что предсказуемость поведения финансового рынка долгое время была и остается предметом дискуссии.

Практическая задача, в рамках цели поставленной для данной работы, состоит в бинарной классификации методом опорных векторов, что реализовано как определение понижения или повышения стоимости индекса Nikkei 225 по данным технических индикаторов. В качестве признаков для классификации данных, т.е. определения к какой категории – повышение цены или понижения, относятся цены были использованы технические индикаторы RSI, EMAcross, MACD, WPR, CCI.

Основным инструментом для решения поставленной задачи является свободное программное обеспечение R, включающее одноименный язык программирования и программную среду вычислений с открытым исходным кодом.

2.1 Сбор и представление данных

Материал для классификации были получены на ресурсе Quandl, где предоставлены финансовые и экономические данные. Информация представляет собой данные о 247 торговых днях в промежуток с 1 июня 2016 до 1 июня 2017 в текстовом формате .csv. Таблицы полученных данных имеет атрибуты «Date» «Open.Price» «High.Price» «Low.Price» «Close.Price», что соответствует сведениям о дате, цене открытия, наивысшей цене за день, самой низкой цене за день и цене закрытия.

Далее, с помощью функций технических индикаторов: RSI(), EMAcross (), MACD (), WPR () и CCI(), вычисляем их значения для каждой строки таблицы.

Для показателя индекса, где цена закрытия больше цены соответствует значение атрибута класс «UP», иначе - «DOWN».

После чего формируем таблицу, где каждая строка содержит метку класса и значения технических индикаторов.

Для реализации метода обучения полученная таблица в соотношении

3:1 делится на обучающую и тестовую выборки.

2.2 Обучение

Реализация для языка R алгоритма обучения машины опорных векторов доступна в пакете e1071.

Для обучения SVM-модели предназначена функция svm, которая для данного случая имеет следующие параметры:

```
SVM <- svm(
  Class ~  . ,
  data   = TrainingSet,
  kernel = "radial",
  type   = "C-classification",
  cost   = 10,
  gamma  = 0.1,
  na.action = na.omit
  )
```

Но прежде чем использовать функцию svm применительно к тренировочной выборке, нужно подобрать оптимальные параметры C и γ для функции ядра. Вариации параметра могут существенно влиять на качество обучения. Для этой цели в пакете e107 реализована функция tune, которая позволяет провести подбор оптимальных параметров с помощью одной функции:

```
tuned <- tune.svm (
                  Class ~ . ,
                  data = DataSet,
                  gamma = 10^(-7:2),
                  cost = 10^(-2:2),
                  na.action = na.omit
                  )
```

Здесь параметры C и γ заданы как промежутки значений, из который нужно выбрать оптимальные. Оптимальные значения этих параметров были получены с помощью 10-кратной перекрестной проверки (кросс-валидации).

2.3 Прогнозирование и оценка корректности

Для осуществления предсказаний на новых данных с помощью обученной SVM-модели в пакете e1071 реализована функция predict():

```
predition <- predict(SVM, TestSet)
```

Как результат он возвращает список, состоящий значений параметра «Class» для каждого элемента тестовой выборки.

Обратимся к способам оценки классификации.

Сначала рассмотрим матрицу ошибок (Confusion matrix).

Матрица неточностей или ошибок представляет собой инструмент, использующий кросс-табуляцию для показа того, как соотносятся значения совпадающих классов, полученные из различных источников.

И представляет следующий результат (1):

```
          actual
predicted DOWN UP
     DOWN   26  1
     UP      5 30
```

Рисунок 1

Следующим инструментом для оценки качества классификации будет кривая ошибок или ROC-кривая.

ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров (**??**).

Количественную интерпретацию ROC дает показатель AUC (area under ROC curve) — площадь, ограниченная ROC-кривой и осью ложных положительных классификаций (2). Чем выше показатель AUC, тем качественнее классификатор.

Эффективность метода опорных векторов зависит от выбора ядра, параметров ядра и параметра C для геометрической разницы.

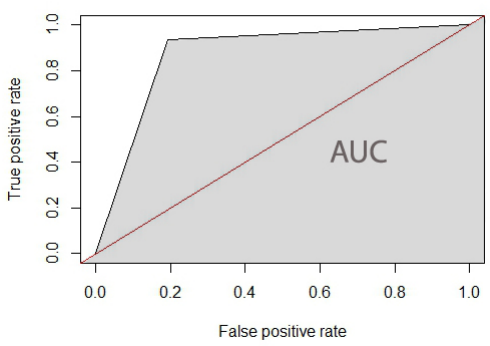

Рисунок 2

ЗАКЛЮЧЕНИЕ

В ходе данной работы был рассмотрен метод опорных векторов для случая бинарной классификации. Поставленные цели были достигнуты путем применения метода опорных векторов для прогнозирования движения фондового индекса Nikkei 225 по данным за год. Были получены результаты, свидетельствующие о достаточной корректности данного метода. Специфика анализа финансового рынка определяется нелинейным характером данных и временных рядов, высокой степенью неопределенности. Линейные методы и большинство сложных моделей нелинейного машинного обучения не могут полностью обеспечить достаточную точность. Влияние на динамику рынка множества не только экономических, но и социальных, политических факторов обуславливает невозможность его полной предсказуемости.

Достоинство метода состоит в том, что для классификации методом опорных векторов, в отличие от большинства других методов, достаточно небольшого набора данных. При правильной работе модели, построенной на тестовом множестве, возможно применение данного метода на реальных данных.

Метод опорных векторов позволяет использовать линейный классификатор для работы с нелинейно разделяемыми данными, сочетая простоту и эффективность. Идея оптимального разделения гиперплоскости приводит к максимизации ширины разделяющей полосы между классами в задачах классификации, что повышает качество обобщения.

Преимуществом использования метода опорных векторов для извлечения сложных паттернов из данных заключается в том, что для его использования нет необходимости предварительно понимать поведение данных. Для анализа данных и извлечения паттернов методу достаточно наблюдений за данными и связями внутри них. Таким образом, метод опорных векторов работает по принципу "черного ящика получая входы и генерируя выходы, которые могут оказаться очень полезными в нахождении паттернов в очень сложных и неочевидных данных.

Одна из лучших особенностей метода опорных векторов заключается в том, что он очень хорошо справляется с ошибками и шумами в данных.