МЕТОД МАШИН ОПОРНЫХ ВЕКТОРОВ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОКАЗАТЕЛЕЙ ИНВЕСТИЦИЙ

УДК 004.942

Ольга Викторовна Китова,
д.э.н., доцент, зав. кафедрой информатики Российского экономического университета имени Г.В. Плеханова (РЭУ им. Г. В. Плеханова)
Тел.: (499) 237 85 20

Игорь Борисович Колмаков,
д.э.н., профессор кафедры информатики Российского экономического университета имени Г.В. Плеханова (РЭУ им. Г. В. Плеханова)
Тел.: (495) 958 24 10

Илья Андреевич Пеньков,
аспирант кафедры информационных технологий Российского экономического университета имени Г.В. Плеханова (РЭУ им. Г. В. Плеханова)
Тел.: (916) 803 53 50

В статье рассматриваются возможности применения интеллектуального метода машинного обучения на основе опорных векторов для прогнозирования показателей инвестиций. Описываются основные преимущества метода над традиционными методами прогнозирования, что позволяет улучшать качество прогноза. Приводятся результаты компьютерных экспериментов по адаптации моделей на основе машин опорных векторов, реализованных с использованием языка программирования Python, к прогнозированию отдельных показателей инвестиций.

Ключевые слова: *машины опорных векторов, машинное обучение, модель прогнозирования, язык программирования Python, показатели инвестиций.*

Olga V. Kitova,
doctor of science (economics), the head of the Academic Department of Informatics, Plekhanov Russian University of Economics (PRUE)
Tel. (499) 237-85-20
E-mail: olga.kitova@mail.ru

Igor B. Kolmakov,
doctor of science (economics), professor lecturer of the Academic Department of Informatics, Plekhanov Russian University of Economics (PRUE)
Tel. (495) 958-24-10
E-mail: kolibor@rambler.ru

Ilya A. Penkov,
postgraduate, the Academic Department of Informatics, Plekhanov Russian University of Economics (PRUE)
Tel. (916) 803-53-50
E-mail: i.job.penkov@gmail.com

SUPPORT VECTOR MACHINE METHOD FOR PREDICTING INVESTMENT MEASURES

Possibilities of applying intelligent machine learning technique based on support vectors for predicting investment measures are considered in the article. The base features of support vector method over traditional econometric techniques for improving the forecast quality are described. Computer modeling results in terms of tuning support vector machine models developed with programming language Python for predicting some investment measures are shown.

Keywords: *support vector machine, machine learning, forecasting model, programming language Python, investment measures.*

1. Введение

Для определения закономерностей, тенденций изменения динамики поведения макроэкономических показателей необходимо комплексное рассмотрения процессов экономического развития в рамках единой модели. Однако многие модели оказываются неприменимыми в условиях неопределенности социально-экономических процессов, усиления неустойчивости механизмов развития экономики особенно в условиях кризисов или политико-экономических ограничений, изменчивости производственно-экономических отношений, нестационарности большинства процессов. Из этого возникает одна из основных проблем прогнозирования показателей инвестиций, поскольку в силу своей природы они оказываются наиболее подверженными инфляционным колебаниям, демонстрируют высокую подвижность и эластичность к изменениям спроса на продукцию и рыночные услуги.

Существующие традиционные методы прогнозирования, базирующиеся на эконометрических принципах моделирования, в условиях нестабильности оказываются неприемлемыми для прогнозирования [1]. На наш взгляд, в современных условиях функционирования социально-экономических систем проблема получения приемлемого прогноза может быть решена за счет комбинирования традиционных классических методов совместно с методами интеллектуального прогнозирования, реализованных на принципах машинного обучения. К таким методам относятся искусственные нейронные сети, генетические алгоритмы, деревья решений, машины опорных векторов, кластеризация и др.

В рамках статьи авторами исследуются возможности метода опорных векторов в виде решения задачи регрессии при прогнозировании показателей инвестиций.

2. Метод машин опорных векторов

Этот класс методов машинного обучения также может использоваться как для решения задач классификации, так и для восстановления регрессии. Машины опорных векторов (SVM – Support Vector Machine) относятся к категории универсальных сетей прямого распространения, как многослойный персептрон и сети на основе радиальных базисных функций.

Идея метода опорных векторов, предложенного Вапником, состоит в построении гиперплоскости, выступающей в качестве поверхности решений, максимально разделяющей положительные и отрицательные примеры из обучающего множества. Более конкретно машина опорных векторов является аппроксимирующей реализацией метода минимизации структурного риска, который основан на том, что уровень ошибок обучаемой машины на тестовом множестве можно представить в виде суммы ошибки обучения и слагаемого, зависящего от измерения Вапника–Червоненкиса. В случае разделяемых множеств получается значение «нуль» для первого слагаемого и минимизируется значение второго слагаемого. Поэтому машина опорных векторов может обеспечить высокое качество обобщения, не обладая априорными знаниями о предметной области конкретной задачи.

В основе построения алгоритма обучения опорных векторов лежит понятие ядра скалярного произведения опорного вектора и вектора, взятого из входного пространства. Опорные векторы представляют собой подмножество обучающей выборки. Различные методы генерации ядра позволяют строить различные обучаемые машины со своими собственными нелинейными поверхностями решений. Например, алгоритм настройки опорных векторов можно использовать для построения следующих основных типов обучаемых машин:

• полиномиальные машины;

• сети на основе радиальных базисных функций;

• двухслойные персептроны (с одним скрытым слоем).

Такая возможность означает, что для каждой из упомянутых сетей можно использовать алгоритм обучения на основе метода настройки опорных векторов, который будет использовать исходный набор обучающего множества для определения количества скрытых элементов [2].

Учитывая основную идею метода опорных векторов, которая заключается в поиске гиперплоскости, позволяющей разделить исходное множество примеров на два класса, важно отметить понятие линейной разделимости образов. Оно связано с возможностью однозначного бинарного разделения примеров множества. При этом в процессе реализации метода происходит поиск образов (примеров), находящихся на границе между двумя классами. Такие образы являются опорными векторами. Эти векторы играют решающую роль в работе обучаемых машин. Они являются теми точками данных, которые лежат ближе всего к поверхности решений (область между границами двух разделяемых классов) и являются самыми сложными для классификации. Они лучше всего указывают на оптимальное размещение поверхности решений.

Модель классификации или регрессии, построенная на основе машин опорных векторов, считается линейно разделимой, если область между границами двух классов оказывается пустой. Это говорит о том, что в результате функционирования машины опорных векторов удалось разделить все примеры исходного множества на два класса.

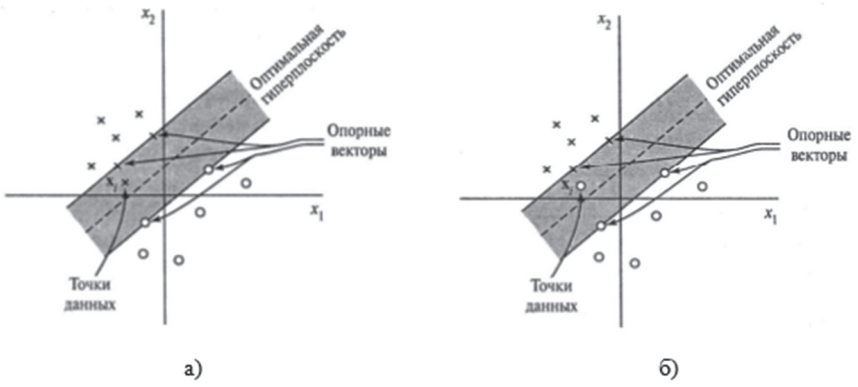

Рис. 1. Варианты попадания точки данных в область разделения

Если исходное множество представляет собой неразделимый набор образов, то при таком наборе данных обучения невозможно построить разделяющую гиперплоскость, полностью исключающую ошибки классификации. В этом случае возможны две ситуации расположения образов в пределах разделяющей гиперплоскости:

• точка находится в области разделения с нужной стороны поверхности решений (если разделить область на две равные части, то точка будет находиться в ближайшей к своему классу половине, рис. 1, а);

• точка попадает в область разделения с другой стороны поверхности решений (рис. 1, б).

В подобной ситуации задача сводится к поиску такой гиперплоскости, которая будет минимизировать ошибку классификации на множестве обучающих примеров.

В общем виде задача классификации при помощи метода машин опорных векторов заключается в поиске некоторой линейной функции, которая разделяет исходное множество образов на два класса и принимает значение меньше нуля для образов (векторов) одного класса и больше нуля – для векторов другого класса.

Наилучшей функцией классификации является функция, для которой ожидаемый риск минимален. Ожидаемый риск – это ожидаемый уровень ошибки классификации.

Напрямую оценить ожидаемый уровень ошибки построенной модели невозможно, это можно сделать при помощи понятия эмпирического риска. Однако следует учитывать, что минимизация последнего не всегда приводит к минимизации ожидаемого риска. Это обстоятельство следует помнить при работе с относительно небольшими наборами обучающих данных.

Эмпирический риск – уровень ошибки классификации на обучающем множестве.

Таким образом, в результате решения задачи методом опорных векторов для линейно разделимых образов мы получаем функцию классификации, которая минимизирует верхнюю оценку ожидаемого риска.

Одной из проблем, связанных с решением задач классификации рассматриваемым методом, является то обстоятельство, что не всегда можно легко найти линейную границу между двумя классами.

В таких случаях один из вариантов – увеличение размерности, т.е. перенос данных из плоскости в трехмерное пространство, где возможно построить такую плоскость, которая идеально разделит множество образов на два класса. Опорными векторами в этом случае будут служить объекты из обоих классов, являющиеся экстремальными. Механизм добавления оператора ядра и дополнительных размерностей позволяет определить границы между классами в виде гиперплоскостей.

Однако следует помнить: сложность построения SVM-модели заключается в том, что чем выше размерность пространства, тем сложнее с ним работать. Один из вариантов работы с данными высокой

размерности – это предварительное применение какого-либо метода понижения размерности данных, например, метод главных компонент или нейронная сеть, реализующая снижение размерности за счет выявления наиболее существенных признаков.

Недостаток метода состоит в том, что для классификации используется не все множество образцов, а лишь их небольшая часть, которая находится на границах.

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, вполне возможно применение данного метода на реальных данных.

Метод опорных векторов позволяет [3]:

• получить функцию классификации с минимальной верхней оценкой ожидаемого риска (уровня ошибки классификации);

• использовать линейный классификатор для работы с нелинейно разделяемыми данными, сочетая простоту с эффективностью.

К основным преимуществам машин опорных векторов обычно относят следующие:

• решается задача квадратичного программирования, имеющая единственное решение, вместо многоэкстремальной задачи;

• происходит автоматическое определение количества скрытых нейронов, которое равно числу опорных векторов;

• оптимальное разделение гиперплоскости приводит к максимизации ширины разделяющей полосы между классами (в задачах классификации), что повышает качество обобщения.

3. Компьютерное моделирование

Компьютерный эксперимент с использованием реализаций модели машин опорных векторов на языке программирования Python проводился для показателей инвестиций в основной капитал по видам экономической деятельности, наблюдаемых Федеральной службой государственной статистики и результаты прогнозов по которым не отвечают заданным критериям точности и качества в эконометрической модели при проведении процедуры ретроверификации [4]. Эти показатели сведены в табл. 1.

Таблица 1

Показатели инвестиций в основной капитал

Обозначение	Показатели инвестиций в основной капитал (ОК) по видам экономической деятельности
I_AHF	Инвестиции в ОК – сельское хозяйство, охота и лесное хозяйство
I_NON	Инвестиции в ОК – добыча полезных ископаемых, кроме топливно-энергетических
I_MME	Инвестиции в ОК – производство машин и оборудования
I_WRT	Инвестиции в ОК – оптовая и розничная торговля
I_HR	Инвестиции в ОК – гостиницы и рестораны

Каждый показатель представляет собой временной ряд длиной 35 наблюдений. Ряд представляет собой квартальные значений показателей. Моделирование проводится на данных до 2013 года включительно, поскольку с 2014 года резко изменилась динамика поведения показателей, что вызвано далеко не экономическими обстоятельствами.

Исходная модель данных, поступающая на вход для каждого показателя, подлежит трансформации следующим образом: учитывая особенности временного ряда, связанные с наличием инерции в процессе, что заключается в актуальности нескольких ближайших наблюдений для формирования последующего значения ряда, исходное множество разбивается на обучающее, включающее первые 75% наблюдений, и тестовое, в которое входят оставшиеся 25% наблюдений.

Прогнозирование на основе машин опорных векторов проводится на основе подхода факторного прогнозирования, когда процесс развития динамики целевой переменной описывается набором определенных макроэкономических переменных, взятых в определенных единицах измерения и форматах (абсолютные значения, доля в ВВП, цепной индекс и т. д).

Мерой сравнения качества получаемых моделей является показатель среднеквадратической ошибки, рассчитываемый на тестовом множестве. Эта мера отражает степень близости значений процесса и его ретропрогноза. Как и для большинства методов машинного обучения, машины опорных векторов требуют проведение предобработки, которая заключается в стандартизации исходных значений ряда (среднее значение = 0, дисперсия = 1).

Модели на основе машин опорных векторов реализованы для случая регрессии и имеют в библиотеку Python Scikit-learn 2 модификации: Epsilon-Support Vector Regression (epsilon-SVR) – модель опорных векторов с ключевым параметром epsilon; Nu Support Vector Regression (Nu-SVR)– модель опорных векторов с ключевым параметром nu.

В реализации epsilon-SVR ключевыми свободно настраиваемыми параметрами являются C – параметр, влияющий на качество построения поверхности решений за счет ее упрощения, epsilon – значение, регулирующее возможность учета ошибки, kernel – функция ядра. Рассмотрим результат компьютерного эксперимента для разных значений параметров. В табл. 2–4 представлены значения среднеквадратической ошибки для модели каждого показателя при разных уровнях параметров C и epsilon и функциями ядра.

Как можно заметить из представленных таблиц, наилучшие результаты по значению среднеквадратической ошибки для каждого показателя демонстрируют модели машин опорных векторов, использующие в качестве параметра ядра (kernel) полиномиальную функцию. Среди моделей, построенных с использованием этой функции, наилучшие результаты при любом значении параметра epsilon демонстрируют реализации, имеющие минимально значение C, равное 1. Для всех остальных значений C модели с поли-

Таблица 2

Результаты расчетов по моделям для радиально-базисной функции ядра

Показатели	Значения C (epsilon = 0,1)				Значения C (epsilon = 0,5)			
	1	10	1000	10000	1	10	1000	10000
I_AHF	18,69	18,61	18,61	18,6	21,08	21,09	21,09	21,09
I_NON	9,63	9,73	9,72	9,72	9,6	9,7	9,7	9,7
I_MME	3,31	3,28	3,28	3,28	3,34	3,29	3,29	3,29
I_WRT	26,42	26,06	26,06	26,06	30,38	30,5	30,5	30,5
I_HR	5,63	6,21	6,21	6,21	5,96	6,11	6,11	6,11

Таблица 3

Результаты расчета по моделям для сигмоидальной функции ядра

Показатели	Значения C (epsilon = 0,1)				Значения C (epsilon = 0,5)			
	1	10	1000	10000	1	10	1000	10000
I_AHF	23,04	23,04	23,04	23,04	23,03	23,03	23,03	23,03
I_NON	9,8	9,8	9,8	9,8	9,74	9,74	9,74	9,74
I_MME	3,82	3,82	3,82	3,82	3,83	3,83	3,83	3,83
I_WRT	33,9	33,9	33,9	33,9	32,92	32,92	32,92	32,92
I_HR	6,27	6,27	6,27	6,27	6,32	6,32	6,32	6,32

Таблица 4

Результаты расчета по моделям для полиномиальной функции ядра

Показатели	Значения C (epsilon = 0,1)				Значения C (epsilon = 0,5)			
	1	10	1000	10000	1	10	1000	10000
I_AHF	12,84	31,58	59,08	68,57	10,53	12,76	86,36	86,36
I_NON	9,25	14,2	52,26	66,05	8,94	12,21	50,15	84,45
I_MME	3,18	2,82	6,27	6,27	2,7	2,61	3,26	4,32
I_WRT	28,21	36,87	63,51	77,13	28,09	31,67	32,26	32,26
I_HR	5,55	9,51	22,4	22,87	5,83	7,9	8,28	8,28

номиальной функцией показывают худшие результаты по сравнению с остальными.

В рамках сравнения моделей с радиально-базисной ядерной функцией для показателей I_AHF и I_WRT преимуществом обладают модели с параметром epsilon=0,1. Для показателей I_NON, I_MME существенной разницы в результатах при разных значениях параметров C и epsilon нет. Для показателя I_HR в модели с радиально-базисной функцией лучшие результаты также получены при значении параметра epsilon=0,1. Однако при фиксированном epsilon наблюдается заметное различие в результатах при разных значениях параметра C. Так, преимущество имеют модели с параметром C=1.

4. Заключение

Построение прогнозных моделей на основе эконометрических методов множественного регрессионного анализа имеет ограничения при прогнозировании временных рядов, имеющих короткую актуальную часть, а также динамика поведения которых определяется не только логикой экономических явлений, что достаточно сильно влияет на характер развития исследуемого процесса. Проведение компьютерного эксперимента показало, что применение метода машин опорных векторов при соответствующем понимании механизмов настройки модели и качестве процедур предобработки исходных данных обладает сильным потенциалом по улучшению параметров точности и качества моделей при решении задачи прогнозирования макроэкономических показателей на основе факторного подхода.

Литература

1. Китова О.В., Дьяконова Л.П., Пеньков И.А. Гибридный подход к прогнозированию показателей инвестиционной сферы // Менеджмент и бизнес-администрирование. 2015. № 3. С. 116–120.

2. Хайкин С. Нейронные сети. Полный курс. – 2-е изд. – М.: Издательский дом «Вильямс», 2006. – 1104 с.

3. B. Scholkopf, G. Ratsch, K. Muller, K. Tsuda, S. Mika An Introduction to Kernel-Based Learning Algorithms // IEEE Neural Networks, 12(2):181–201, May 2001.

4. Колмаков И.Б., Потапов С.В., Пеньков И.А. Верификатор системы прогноза показателей инвестиций экономики РФ // Свидетельство о государственной регистрации программы для ЭВМ № 2016613913, 11.04.2016

References

1. Kitova O.V., Djakonova L.P., Penkov I.A. Hybrid approach to forecasting investment measures // Menedzhment i biznes-administrirovanie. 2015. № 3. p. 116–120.

2. Haykin S. Neural Networks. A comprehensive foundation. – second edition, Prentice Hall, 1999

3. B. Scholkopf, G. Ratsch, K. Muller, K. Tsuda, S. Mika An Introduction to Kernel-Based Learning Algorithms // IEEE Neural Networks, 12(2):181–201, May 2001.

4. Kolmakov I.B., Potapov S.V., Penkov I.A. The investment measures forecasting verifier system in the economy of the Russian Federation // Svidetel'stvo o gosudarstvennoj registracii programmy dlja JeVM № 2016613913, 11.04.2016.