УДК 004.9
DOI 10.26118/2782-4586.2024.87.62.004

ЛАРКИНА АЛЛА АНАТОЛЬЕВНА
Самарский государственный экономический университет
ПОРТНОВ КОНСТАНТИН ВАЛЕРЬЯНОВИЧ
Самарский государственный технический университет
АНИКИН ДМИТРИЙ ВАЛЕРЬЕВИЧ
Самарский государственный технический университет

Алгоритм формирования обучающей выборки на основе метода кластеризации

Аннотация.В статье рассматриваются актуальные аспекты, связанные с процессом формирования обучающей выборки данных, который является ключевым элементом в процессе обучения искусственной нейронной сети, предназначенной для функционирования в качестве биржевого торгового БОТа. Мы постараемся всесторонне осветить все тонкости и нюансы, которые необходимо учитывать при подготовке этой выборки, чтобы обеспечить максимальную эффективность и точность в работе торгового алгоритма. Это позволит нам достичь наилучших результатов в автоматизации торговых операций на бирже и повысить общую продуктивность торговых стратегий, реализуемых с помощью искусственного интеллекта. Предложен метод, направленный на кластеризацию данных обучающей выборки. Этот метод позволяет эффективно группировать схожие по определенным критериям данные, что, в свою очередь, способствует повышению точности и эффективности обучающих алгоритмов.

Ключевые слова: торговый бот, обучающая выборка, искусственная нейронная сеть, кластеризация.

LARKINA ALLA ANATOLYEVNA
Samara State University of Economics
PORTNOV KONSTANTIN VALERIANOVICH
Samara State Technical University
ANIKIN DMITRY VALERIEVICH
Samara State Technical University

Algorithm for forming a training sample based on the clustering method

Abstract. The article discusses the current aspects related to the process of forming a training data sample, which is a key element in the training process of an artificial neural network designed to function as a stock exchange trading BOT. We will try to comprehensively cover all the subtleties and nuances that must be taken into account when preparing this sample in order to ensure maximum efficiency and accuracy in the operation of the trading algorithm. This will allow us to achieve the best results in automating trading operations on the exchange and increase the overall productivity of trading strategies implemented using artificial intelligence. A method is proposed aimed at clustering the training sample data. This method allows you to effectively group similar data according to certain criteria, which, in turn, helps to increase the accuracy and efficiency of training algorithms.

Key words: trading bot, training sample, artificial neural network, clustering.

Введение

Ранее авторами предлагались алгоритмы торговых биржевых ботов, основные задачи которых является определение инвестиционных инструментов, объёмов позиций, а также нахождения наилучших точек открытия инвестиционных позиций. Учитывая рыночную волатильность и постоянно меняющиеся характеристики инвестиционных активов, предлагается методология формирование обучающих выборов с использованием кластерного анализа.

Постановка задачи

В настоящее время в алгоритмах торговых БОТов широко используются модели искусственного интеллекта - искусственные нейронные сети (ИНС). Они позволяют решать слабо формализованные задачи. Несмотря на большие преимущества, представляемые аппаратом ИНС, их использование связано с рядом проблем. Наибольшую сложность представляет собой формирование обучающей выборки. Рыночные активы и среда в которой они торгуются,

представляют собой объект с высоким уровнем изменчивости параметров как среды так и характеристик инструментов и наличие одинаковых условий в разные моменты времени мало вероятно. Таким образом, процесс обучения ИНС требует постоянного обновления данных, что делает задачу сбора и анализа информации не только трудоемкой, но и требующей значительных временных затрат. К тому же, необходимо учитывать, что финансовые рынки подвержены влиянию множества внешних факторов, таких как экономические новости, политические события и даже социальные сети, что еще больше усложняет процесс создания адекватной обучающей выборки. Поэтому, несмотря на потенциал ИНС в торговле, их эффективность напрямую зависит от качества и актуальности входных данных.

Алгоритмы торговых БОТов могут быть как полностью автоматизированными, так и полуавтоматизированными, когда трейдер принимает окончательные решения на основе сигналов, генерируемых системой, в обоих случаях основной задачей алгоритма является распознавание сигналов и отличие их от ложных сигналов.

Решение задачи

Предлагаемый алгоритм кластеризации представляет собой статистическую процедуру выделения групп из имеющегося набора данных. Преимуществом этого метода является то, что он по некоторым признакам выявляет схожесть или различие выборок, не требуя при этом вмешательства экспертов и априорной информации о классах.

Кластеризация обучающих выборок — это метод анализа данных, который используется для группировки объектов (в данном случае, обучающих выборок) в кластеры на основе их сходства. Это позволяет выделить естественные структуры в данных, что может быть полезно для различных задач, таких как классификация, сегментация или выявление аномалий. Кластеры представляют собой группы объектов, которые имеют высокую степень схожести внутри группы и низкую степень схожести с объектами из других групп. Это может быть основано на различных

характеристиках объектов, таких как признаки, значения или метрики.

Предлагаемый геометрический метод распознавания основан на использовании функции расстояния в качестве меры сходства векторов признаков, представляющих образцы. В качестве меры близости (или подобия) двух точек используется евклидово расстояние между ними.

Метод кластеризации не учитывает специфику исследуемых объектов, следовательно, был разработан алгоритм, который позволяет формировать обучающую выборку на множестве исходных данных, для определения наилучших решений относительно позиции(набор активов и моменты входа и выхода в рынок) торгового БОТа.

Процедура вычислений предусматривает выполнение следующих шагов.

Шаг 1. Анализируем важность всех параметров и исключаем из них все малозначащие и те, которые мало отличаются у сравниваемых объектов.

Шаг 2. Формируем значения исходных векторов в виде двумерной матрицы p .

Шаг 3. Устанавливаем количество кластеров к.

Шаг 4. Задаем значения центров кластеров в виде двумерной матрицы m .

Шаг 5. Находим расстояние между центрами кластеров и всеми векторами по формуле 1:

$$d_{ti} = \sum_{j=1}^{n} (p_{ti} - m_{ij})^2 \quad (1)$$

где $t \in [1, n]$, где n - количество исходных векторов; где $i \in [1, k]$, где k – количество кластеров; где $j \in [1, r]$, где r - размерность входного вектора; где $z \in [1, zn]$, где zn - количество векторов, принадлежащих i-му классу.

Шаг 6. Определяем принадлежность векторов к кластерам по формуле 2

$$Index(p_{tj})=min(d_{ti}) \quad (2)$$

Шаг 7. Сохраняем центры кластеров, по формулам 3,4:

$$m_{copy}:=m \quad (3)$$

$$^{m}cop_{Y} := {}^{m} \quad (4)$$

Шаг 8. Определяем новые значения центров кластеров, по формуле 5

$$m_{ij} = \frac{1}{nz}\sum_{z=1}^{nz} P_{zj} \quad (5)$$

Шаг 9. Если $m_{copy} \approx m$, то переходим на Шаг 10, если нет - возвращаемся на Шаг 5.

Шаг 10. Проверяем принадлежность прогнозируемого образца к одному из кластеров по формуле (1).

Тот кластер, к которому принадлежит исследуемый образец, составляет обучающую выборку, а параметры которые описывали объект, могут использоваться как входы ИНС.

По данной модели было разработано программное обеспечение (ПО), которое упрощает задачу расчета промежуточных данных, и позволяет графически просмотреть расположение кластеров, их центров, и принадлежащих им образцов в 3-х мерном пространстве.

Результаты исследования проводились, для простейшего случая кластеризации обучающих выборок в двухмерном пространстве риск-доходность. Риска инвестиционного актива в общем виде оценивается размером потенциальных потерь, которые могут возникнуть при инвестировании в конкретный финансовый инструмент, такой как акции, облигации или другие ценные бумаги. Существует несколько подходов к оценке данного показателя. Доходность в данном случае определяется размером потенциальных денежных поступлений в будущем.

Пространство поиска изображено на рисунке 1.

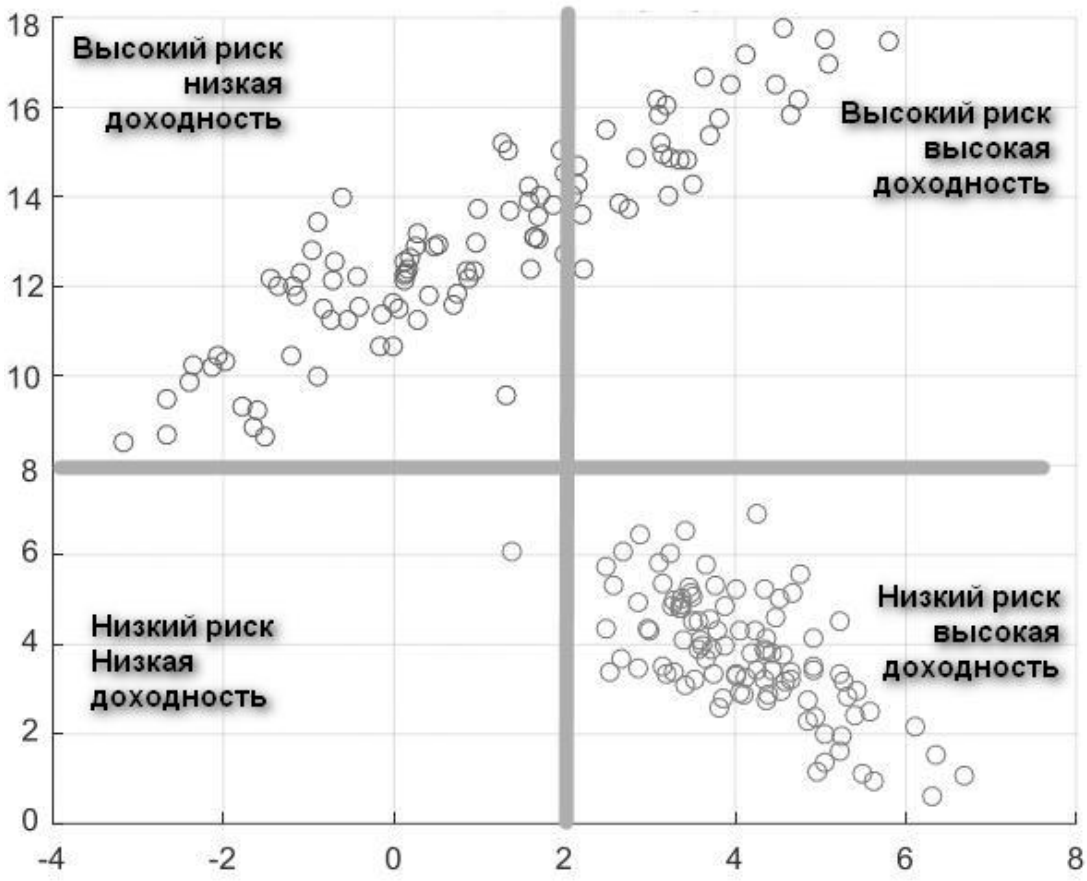

Рисунок 1 – Пример целевого кластера

Очевидно что целевым кластером будет являться правой нижней кластер с минимальным риск на максимальной доходностью. Существуют разные подходы к оценке доходности и риска но в нашей статье, это не принципиально и мы не будем касаться данного вопроса. В зависимость между риском и доходом всегда однозначна - чем больше риск тем больше доходность.

Для повышения качества обучения ИНС необходимо устранить из выборки ошибочные значения представляющая собой статистическую погрешность.

Выводы

Анализ классов в рамках обучающей выборки с помощью кластерного подхода служит ключевым инструментом для формирования надежных систем классификации. Использование гистограммы и соответствующий ей метод уменьшения размерности кластеров эффективно способствует

выявлению врожденной кластерной организации данных. После кластеризации метод дополнения кластеров точками выборки способствует максимально полному использованию обучающих материалов. На основе структурированности кластеров удобно формировать стандарты для каждого класса, что является основой для создания результативной системы классификации. Методика оценки формы кластеров способствует повышению точности их разделения. Результатом использования этой методики является модель классификатора, где каждому классу соответствует определенный набор кластеров, а каждому кластеру — соответствующий набор стандартов. Для создания таких классификаторов можно применять эффективные нейронные сети с глубоким обучением, которые обучаются с использованием эталонных образцов классов.

Список использованных источников:

1. Олин Р. А. Влияние автономных агентов с искусственным интеллектом на инновационные процессы в корпоративной среде / Р. А. Олин // Актуальные тренды в развитии науки, экономики, образования : Сборник научных статей Всероссийской научно-практической конференции, Самара, 17 июня 2024 года. – Самара: Самарский государственный экономический университет, 2024. – С. 18-23. – EDN KFZBZR.

2. Олин Р. А. Формирование инновационной среды предприятия с использованием средств искусственного интеллекта / Р. А. Олин, Е. С. Шатрова // Журнал монетарной экономики и менеджмента. – 2024. – № 3. – С. 213-217. – DOI 10.26118/2782-4586.2024.99.23.032. – EDN EMVQGB.

3. Олин Р. А. Метавселенные как часть цифровой трансформации / Р. А. Олин // Устойчивое развитие в неустойчивом мире : Сборник научных статей Международной научно-практической конференции, Самара, 23 мая 2023 года. – Самара: Самарский государственный экономический университет, 2023. – С. 297-306. – DOI 10.46554/UR-2023-pp.297. – EDN USQROQ.

4. Олин Р. А. Инновации и управление рисками с использованием NFT / Р. А. Олин // Наука XXI века: актуальные направления развития. – 2023. – № 2-2. – С. 292-298. – DOI 10.46554/ScienceXXI-2023.09-2.2-pp.292. – EDN YCZSUI.

5. Олин Р. А. Информационное общество: сущность и проблемы становления в Российской Федерации / Р. А. Олин // Современные проблемы управления : Сборник научных статей / Под редакцией С.А. Ключникова, Самарский государственный университет», факультет экономики и управления, кафедра государственного и муниципального управления. Том Выпуск 6. – Самара : Издательство «Глагол», 2013. – С. 178-183. – EDN WXVDGJ.

6. Панюков Д. И. Информационные технологии поддержки систем менеджмента качества в автопроме / Д. И. Панюков, Е. В. Панюкова // Синергетика природных, технических и социально-экономических систем. – 2018. – № 15. – С. 192-198. – EDN HMNEEC.

7. Панюкова Е. В. Использование компетентностного подхода в преподавании математики и информатики / Е. В. Панюкова, Э. В. Егорова // Проблемы университетского образования. Компетентностный подход в образовании : сборник материалов IV Всероссийской научно-методической конференции: в 3 томах, Тольятти, 10–11 декабря 2009 года / Тольяттинский государственный университет. Том I. – Тольятти: Тольяттинский государственный университет, 2009. – С. 269-274. – EDN VDLHMP.

8. Панюков Д. И. Повышение статуса инженерно-технического образования в современных условиях / Д. И. Панюков, Е. В. Панюкова // Синергетика природных, технических и социально-экономических систем. – 2017. – № 14. – С. 204-208. – EDN ZFTUPP.

9. Панюкова Е.В Автоматизация мониторинга как новая форма контроля учебных достижений студентов и управления образовательной деятельностью / О. М. Гущина, Е. В. Панюкова, О. В. Аникина // Азимут

научных исследований: педагогика и психология. – 2017. – Т. 6, № 3(20). – С. 72-75. – EDN ZISRIN.

10. Панюкова Е.В. Система анализа цифровых контентов в образовательной среде / Д. И. Панюков, Е. В. Панюкова // Синергетика природных, технических и социально-экономических систем. – 2018. – № 15. – С. 178-181. – EDN YAUWPR.

11. Латушкина Т. С. Анализ проблем квалиметрии профессиональных знаний / Т. С. Латушкина, Н. Ю. Портнова, Е. В. Сибарцева // Вызовы современности и стратегии развития общества в условиях новой реальности (шифр -МКВСС) : Сборник материалов XXVII Международной научно-практической конференции, Москва, 10 июня 2024 года. – Москва: ООО "Издательство "Экономическое образование", 2024. – С. 385-391. – DOI 10.26118/6979.2024.94.21.002. – EDN KATYSK.

12. Латушкина Т. С. Анализ подходов используемых в квалиметрии профессиональных знаний / Т. С. Латушкина, Н. Ю. Портнова, Е. В. Сибарцева // Вызовы современности и стратегии развития общества в условиях новой реальности (шифр -МКВСС) : Сборник материалов XXVII Международной научно-практической конференции, Москва, 10 июня 2024 года. – Москва: ООО "Издательство "Экономическое образование", 2024. – С. 375-384. – DOI 10.26118/8575.2024.28.11.003. – EDN IRTYSH.

Информация об авторах

ЛАРКИНА АЛЛА АНАТОЛЬЕВНА, к.э.н., доцент каф. прикладная информатика, ФГАОУ ВО «Самарский государственный экономический университет», г. Самара, Россия

ПОРТНОВ КОНСТАНТИН ВАЛЕРЬЯНОВИЧ, к.т.н., доцент кафедры «ИВТ», ФГБОУ ВО «Самарский государственный технический университет», г. Самара, Россия

АНИКИН ДМИТРИЙ ВАЛЕРЬЕВИЧ, ассистент, каф."ИВТ", ФГБОУ ВО «Самарский государственный технический университет», г. Самара, Россия

Information about the authors

LARKINA ALLA ANATOLYEVNA, PhD, Associate Professor, Department of Applied Informatics, Samara State University of Economics, Samara Russia

PORTNOV KONSTANTIN VALERIANOVICH, PhD, Associate Professor, Department of ICT, Samara State Technical University, Samara, Russia

ANIKIN DMITRY VALERIEVICH, Assistant, Department of ICT, Samara State Technical University, Samara, Russia.