УДК 330.4+004.8

Методы интеллектуального анализа данных для прогнозирования финансовых временных рядов

Г. Ю. Чернышова, Е. А. Самаркина

Чернышова Галина Юрьевна, кандидат экономических наук, доцент кафедры дискретной математики и информационных технологий, Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, galacherny@yandex.ru

Самаркина Екатерина Александровна, студентка магистратуры, Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, sea94@inbox.ru

Введение. Совершенствование алгоритмов интеллектуального анализа данных позволяет решать задачи прогнозной аналитики более эффективными способами. Ансамбли моделей – одно из активно развивающихся направлений, особенно в тех задачах, где прогностическая точность более важная, чем интерпретируемость модели. **Теоретический анализ.** Задача прогнозирования требует тщательного исследования исходного набора данных и методов, подходящих для анализа. Она включает решение таких подзадач, как выбор модели прогнозирования, анализ точности построенного прогноза. Ансамблевые модели позволяют сочетать прогнозы нескольких базовых моделей с целью уменьшения ошибок прогнозирования и повышения обобщающей способности моделей. Для основных методов построения ансамблей (бэггинг, бустинг, стекинг) представлены концептуальные схемы. **Эмпирический анализ.** В статье исследованы практические аспекты прогнозирования финансовых временных рядов, реализован набор ансамблевых моделей для решения прогностических задач. **Результаты.** Для построения ансамблей моделей разработано приложение с web-интерфейсом, которое позволяет оценить модели с применением различных функционалов ошибки, осуществить выбор более точной ансамблевой модели. С помощью приложения можно осуществлять выбор и настройку моделей для формирования прогнозных ансамблей, тестировать полученные модели, визуализировать результаты. Представлено тестирование ансамблевых моделей на реальных данных для анализа финансовых временны́х рядов.

Ключевые слова: прогнозирование, финансовые временные ряды, интеллектуальный анализ данных, ансамбли моделей.

DOI: https://doi.org/10.18500/1994-2540-2019-19-2-181-188

Введение

Определяющим трендом использования финансовых инструментов является минимизация возможных рисков. Тенденция снижения волатильности, отмечаемая с 2015 г., в настоящее время изменилась, и наблюдается высокая изменчивость на российском финансовом рынке. Рост волатильности связан с целым комплексом взаимосвязанных факторов, в частности, с уменьшением объемов денежных средств, размещаемых в финансовые инструменты, изменениями мировых цен на нефть, колебаниями курса национальной валюты, сохранением геополитической напряженности и мировых экономических диспропорций. В подобной ситуации для участников фондового рынка важной задачей становится применение эффективных способов анализа и прогнозирования ценовых колебаний с целью планирования рисков.

Математические модели и методы становятся важным инструментом обеспечения процесса принятия решений в данной области. Прогнозирование нестационарных процессов является существенным элементом инвестиционной деятельности. Исследование динамики финансовых временных рядов позволяет принять правильное решение об инвестициях. Рыночный механизм, характеризующийся огромным количеством постоянно меняющихся связей, зависит от множества внешних факторов, способных существенно повлиять на всю структуру его зависимостей, причем воздействие может быть самым разнообразным. Появление тех или иных внешних факторов не всегда отражается в предыстории финансового временного ряда, но может вызвать значительное нарушение его динамики.

Для прогнозирования финансовых временных рядов применяются различные подходы:

– графические методы (графики японских свечей);

– экспертные методы (метод Дельфи, сценарные методы);

– стохастические методы (модель броуновского движения, скользящие средние, модель Брауна, Хольта–Винтерса);

– эконометрические модели (регрессионный анализ, авторегрессионные модели, ARIMAX, GARCH, ARDLM);

– методы имитационного моделирования;

– методы Data Mining (нейронные сети, деревья решений, метод ближайшего соседа, метод опорных векторов, ансамбли моделей);

– модели нечеткой логики;

– генетические алгоритмы;

– методы нелинейной динамики;

– методы фрактального анализа;

– модели на основе вейвлет-преобразования.

© Чернышова Г. Ю., Самаркина Е. А., 2019

Следует отметить возможность применения гибридных моделей для решения прогностических задач. Математические модели, использующие методы информационного поиска и средства нечеткой логики, позволяют использовать элементы как технического, так и фундаментального анализа [1].

Широкое применение методов интеллектуального анализа в данной области обусловлено наличием в финансовых временных рядах сложных закономерностей, которые не обнаруживаются линейными методами. Нестационарность этих рядов требует специализированных моделей и методов для прогноза.

Прогнозирование определяет наиболее общие показатели перспективного развития, выявляет тенденции и альтернативные пути этого развития. Методы математической статистики используются, главным образом, для проверки заранее сформулированных гипотез, а также для анализа, составляющего основу оперативной аналитической обработки данных, что является недостаточным условием для полной оценки прогнозирования. Для обработки слабоструктурированных данных необходимо применение специфических методов интеллектуального анализа данных.

Основной проблемой при решении задачи прогнозирования является получение разумно точных прогнозов будущих данных при анализе имеющейся информации. Если методы интеллектуального анализа данных не обеспечивают достаточной точности прогнозирования, эффективной альтернативой использованию методов прогнозирования является внедрение ансамблей моделей [2]. Целью исследования выступает разработка и апробация ансамблевых моделей для прогнозирования рынка ценных бумаг.

Теоретический анализ

В рамках интеллектуального анализа данных выделяются следующие задачи: классификация, прогнозирование, кластеризация, визуализация, поиск ассоциативных правил. По назначению они делятся на описательные и предсказательные задачи. По способам решения задачи разделяют на обучение с учителем и обучение без учителя [3].

Под прогнозированием обычно понимается моделирование непрерывных значений, в отличие от задачи классификации, связанной с получением дискретных прогнозов.

Набор данных обычно состоит из векторов признаков, где каждый вектор представляет собой описание объекта с использованием набора показателей.

В Data Mining категориальные и числовые величины обрабатываются при помощи соответствующих алгоритмов.

Задачи классификации и прогнозирования сводятся к определению значения зависимой переменной объекта по его независимым переменным. Если зависимая переменная принимает количественные значения, то говорят о задаче прогнозирования, в противном случае – о задаче классификации. Задача прогнозирования может считаться одной из наиболее сложных задач Data Mining, она требует тщательного исследования исходного набора данных и методов, подходящих для анализа.

При создании алгоритмов машинного обучения разработчики сталкиваются с такими проблемами, как вычислительные затраты на реализацию алгоритма, неясность построенных моделей для пользователя, а также неточность результатов. Большинство исследователей сосредоточиваются на повышении точности прогнозирования, поэтому оценка моделей часто рассматривается именно с этой точки зрения.

Эффективной альтернативой использованию единственного метода прогнозирования является объединение прогнозов из нескольких разных моделей. Комбинация нескольких прогнозов в большом количестве случаев существенно снижает общие ошибки прогнозирования, превосходя отдельные компоненты [3].

Ансамблевые методы представляют собой метаалгоритмы, которые объединяют несколько алгоритмов в одну прогностическую модель для повышения точности прогноза, уменьшения дисперсии или смещения, позволяя строить более стабильные и надежные модели [4].

Можно выделить два подхода при построении ансамблевых моделей:

– последовательное применение обучающих алгоритмов;

– параллельное использование обучающих алгоритмов.

При этом может использоваться единый базовый алгоритм либо применяются базовые алгоритмы разного типа.

В настоящее время разработано множество различных подходов к формированию ансамблей [5]. Среди них наибольшее распространение получили такие методы, как бэггинг, бустинг, стекинг и блендинг.

Бэггинг (bootstrap aggregating) представляет собой один из способов уменьшить дисперсию результатов прогнозирования путем усреднения нескольких оценок. В этом случае исходная выборка разделяется на подмножества для обучения базовых алгоритмов (рис. 1). Для агрегирования результатов базовых моделей в прогнозных за-

дачах используется усреднение с использованием среднего арифметического значения или средневзвешенного значения.

Бустинг (boosting) относится к семейству алгоритмов, которые способны использовать слабые модели, которые лишь немногим лучше, чем случайные предположения. Принципиальное отличие бустинга заключается в том, что базовые алгоритмы обучаются последовательно на основе изменяемой выборки (рис. 2).

Стекинг (stacking) – это метод обучения ансамбля, который объединяет несколько прогнозных моделей с помощью метаалгоритма. Модели базового уровня формируются на основе полного обучающего набора, затем метамодель строится на выходах моделей базового уровня (рис. 3).

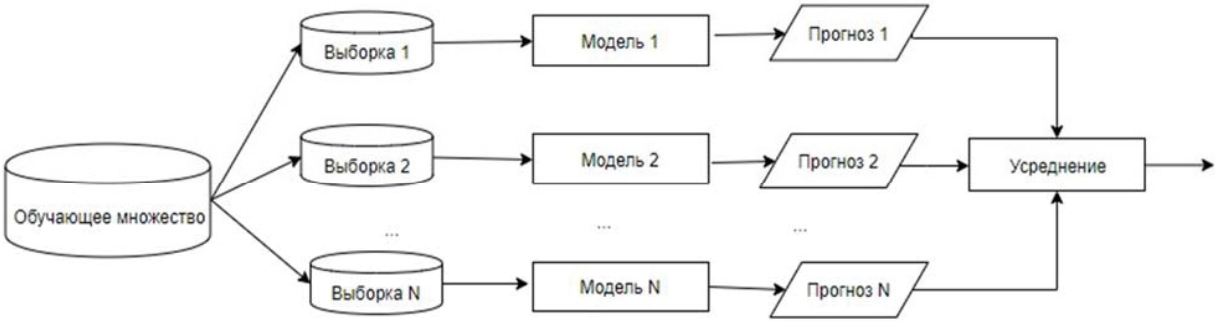

Рис. 1. Построение ансамблей методом бэггинга
Fig. 1. Concept Diagram of Bootstrap Aggregating

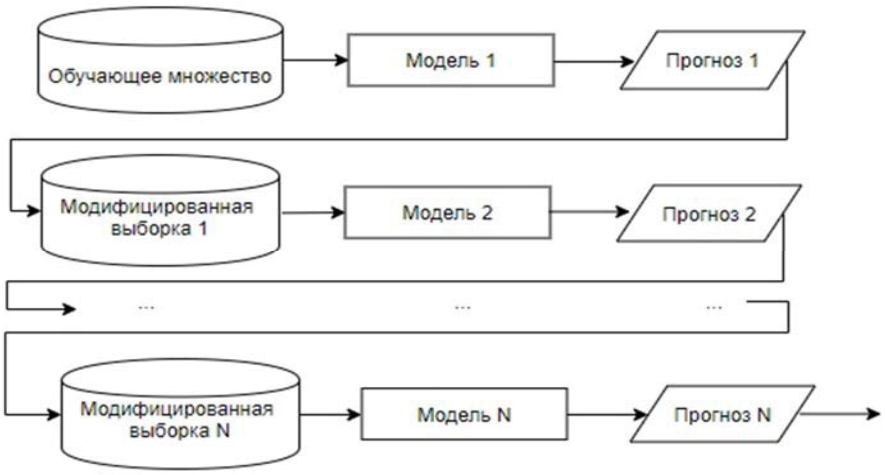

Рис. 2. Построение ансамблей методом бустинга
Fig. 2. Concept Diagram of Boosting

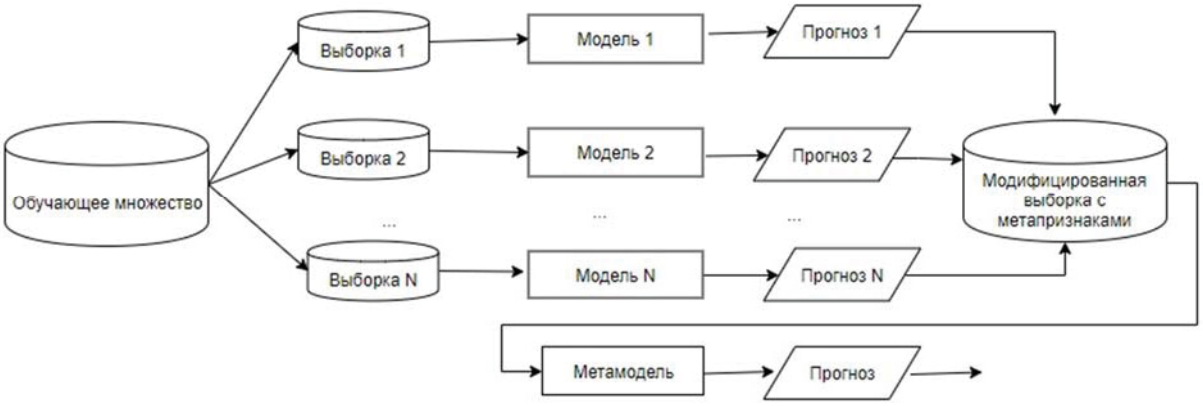

Рис. 3. Построение ансамблей методом стекинга
Fig. 3. Concept Diagram of Stacking

Базовый уровень часто состоит из алгоритмов разного типа, и поэтому стековые ансамбли часто неоднородны.

Простейшая схема стекинга – блендинг (blending). В этом случае обучающую выборку делят на две части. На первой обучают базовые алгоритмы. Результаты прогнозирования с помощью базовых алгоритмов можно рассматривать как новый признак (метапризнак). На метапризнаках второй части обучения настраивают метаалгоритм. Затем запускают его на метапризнаках тестового набора и получают окончательный прогноз. Блендинг и стекинг являются похожими подходами, основаными на обучении первого слоя моделей и использовании их выходов для обучения второго уровня моделей. При этом можно использовать любое количество слоев.

Для оценки моделей прогнозирования используются разные метрики качества, определяющие различие между полученным в результате применения модели значением и реальным значением. Пусть имеется выборка $X = (X_1, \ldots, X_n)$, на которой оценивается прогнозная модель M, $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)$, где $\hat{Y}_i = M(X_i)$, $I = 1, \ldots, n$ – вектор прогнозов, $Y = (Y_1, \ldots, Y_n)$ – вектор наблюдаемых значений прогнозируемой переменной, $Y_{\max}$ – максимальное значение Y_i, $Y_{\min}$ – минимальное значение Y_i, $\bar{Y}$ – среднее значение для $Y_1, \ldots, Y_n$.

Расширенный набор функций ошибки представлен в табл. 1.

Таблица 1/Table 1

Функции ошибки для оценки модели прогнозирования
Error Metrics for Forecasting Model Evaluation

Название	Формула	Краткое описание
Коэффициент детерминации (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n} \lvert Y_i - \hat{Y}_i \rvert^2}{\sum_{i=1}^{n} \lvert \bar{Y} - Y_i \rvert^2}$	Доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью
MSE (Mean Squared Error)	$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	Среднеквадратичная ошибка
RMSE (Root-Mean-Square Error)	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$	Квадратный корень из среднеквадратичной ошибки
NRMSE (Normalized Root-Mean-Square Error)	$\mathrm{NRMSE} = \frac{RMSE}{Y_{max} - Y_{min}}$	Нормализованная RMSE
RAE (Relative Absolute Error)	$\mathrm{RAE} = \frac{\sum_{i=1}^{n} \lvert Y_i - \hat{Y}_i \rvert}{\sum_{i=1}^{n} \lvert \bar{Y} - Y_i \rvert}$	Отношение суммы модулей отклонений к суммарной ошибке
RRSE (Root Relative Squared Error)	$\mathrm{RRSE} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$	Относительная среднеквадратичная ошибка
MAPE (Mean Absolute Percentage Error)	$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \frac{\lvert Y_i - \hat{Y}_i \rvert}{\lvert Y_i \rvert}$	Средний процент отклонения
SMAPE (Symmetric Mean Absolute Percentage Error)	$\mathrm{SMAPE} = \frac{2}{n}\sum_{i=1}^{n} \frac{\lvert Y_i - \hat{Y}_i \rvert}{\lvert Y_i + \hat{Y}_i \rvert}$	Симметричный средний процент отклонения

Всегда существует вероятность события, которое не может предсказать модель, создающая прогноз (событие «черного лебедя»). Ошибки, связанные с этими событиями, не являются типичными, которые рассмотренные функционалы ошибки пытаются измерить. Таким образом, существуют ограничения при использовании моделей для прогнозирования будущего.

Эмпирический анализ

Применение ансамблей моделей связано с проблемой подбора параметров как в процессе выбора типа ансамбля, так и при выборе и настройке параметров базовых алгоритмов [6]. Возникает необходимость использования дополнительных инструментальных средств для решения подобной задачи на основе минимизации функционалов ошибки.

Разработанное web-приложение на платформе Python реализует следующие функциональные возможности: загрузка файла с входными данными; выбор вида ансамблевой моделей; задание параметров ансамбля; выбор метода прогнозирования; задание параметров для реализации методов прогнозирования; прогнозирование на основе выбранной модели, визуализация результатов прогнозирования на основе построенной модели.

Для выбранного метода анализируется точность прогноза при увеличении размера ансамбля. Для используемых методов рассчитываются ранги с учетом выбранных функций ошибки. В результате выбирается модель, имеющая наилучшие ранги по большему количеству различных функций ошибки, с учетом приоритета нормализованной RMSE.

С помощью данного приложения выполнено построение прогнозной модели для акций компаний ООО «Яндекс» и Mail.ru Group. Для обучения и тестирования моделей использованы данные о цене акций компаний за период 2010–2018 гг. [7]. Для формирования обучающего множества применен метод скользящего окна. В качестве базовых методов использовались различные методы для решения задачи прогнозирования (метод опорных векторов, метод ближайшего соседа, деревья решений). С помощью разработанного приложения осуществлен подбор параметров алгоритмов и выбран лучший алгоритм на основе вычисленных значений функционалов ошибки.

При использовании сложных моделей необходимо оценить обобщающую способность используемых методов, для эмпирического измерения явления переобучения модели применяется скользящий контроль. Тестовые выборки для оценки моделей у каждого из ансамблей могут иметь различный размер, зависящий от предполагаемого преобразования исходной выборки данных.

В табл. 2 содержится сравнение функционалов ошибки для отдельного метода и построенных ансамблей моделей для прогнозирования цены акций ООО «Яндекс». В качестве базового алгоритма выбран метод опорных векторов (SVM).

Таблица 2/Table 2

Оценка ансамблей моделей для прогнозирования цены акций ООО «Яндекс»
Evaluation of Models Ensembling for Predicting Yandex LLC Share Price

Ансамблевая модель	R^2	MSE	RMSE	NRMSE	RAE	RRSE	MAPE	SMAPE
Бэггинг	0.847	5.330	2.309	0.111	0.308	0.443	0.051	0.050
Бустинг	0.842	6.951	2.637	0.139	0.309	0.491	0.052	0.05
Блендинг	**0.939**	2.709	1.646	**0.070**	**0.170**	**0.278**	0.032	0.032
Стекинг	0.917	**2.572**	**1.604**	0.074	0.230	0.333	**0.029**	**0.028**
SVM	0.830	6.172	2.484	0.116	0.236	0.437	0.041	0.038

Для указанного набора данных блендинг является предпочтительной моделью с учетом ранжированного сравнения ошибок. В данном случае при использовании блендинга тестовая выборка составляла 600 дней.

Для визуализации результатов прогнозирования приложение позволяет построить графики, обеспечивающие возможность пользователю в процессе оценки модели получить дополнительную информацию о качестве построенной модели. На рис. 4 представлены графики для сравнения результатов прогнозирования с помощью блендинга цены акций ООО «Яндекс» и значений из тестовой выборки.

В табл. 3 представлены вычисленные значения функционалов ошибки для построенных в приложении ансамблей моделей для акций компании Mail.ru Group.

Для этого набора данных применение стекинга позволяет построить более точную модель с учетом вычисленных значений ошибок. В данном ансамбле в качестве базовых алгоритмов использовались метод опорных векторов и метод ближайших соседей, в качестве метаалгоритма на втором этапе − метод ближайших соседей.

На рис. 5 представлены прогнозные значения, полученные в результате использования

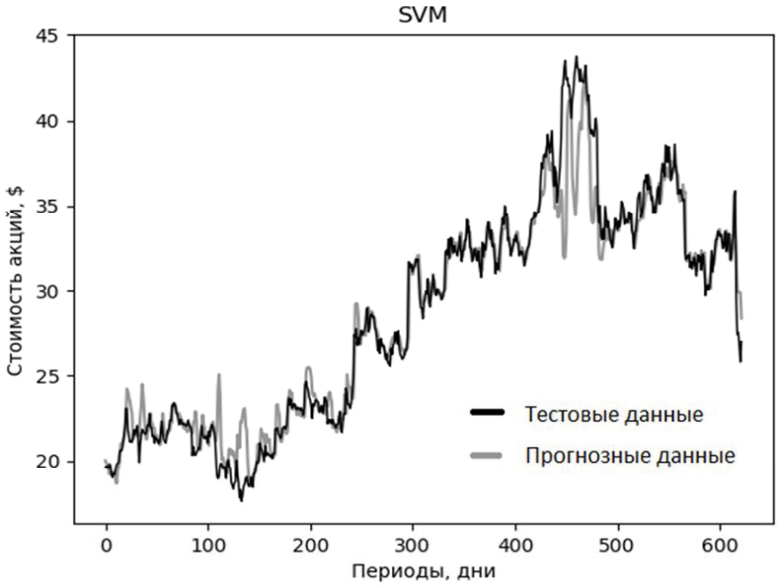

Рис. 4. Сравнение прогнозных и фактических цен акций ООО «Яндекс» методом блендинга на тестовой выборке

Fig. 4. Comparison of Forecast and Actual Yandex LLC Share Prices by Blending for the Test Sample

Таблица 3/Table 3

Оценка ансамблей моделей для прогнозирования цены акций Mail.ru Group

Evaluation of Models Ensembling for Predicting Mail.ru Group Share Price

Ансамблевая модель	R^2	MSE	RMSE	NRMSE	RAE	RRSE	MAPE	SMAPE
Бэггинг	0.411	6.694	2.587	0.174	0.665	0.709	0.079	0.084
Бустинг	0.803	6.351	2.52	0.165	0.571	0.671	0.073	0.078
Блендинг	**0.945**	1.819	1.349	0.072	0.160	0.259	0.026	0.026
Стекинг	0.906	**1.311**	**1.145**	0.081	0.252	0.352	**0.023**	**0.022**
SVM	0.841	1.326	1.152	**0.057**	**0.156**	**0.214**	0.026	0.026

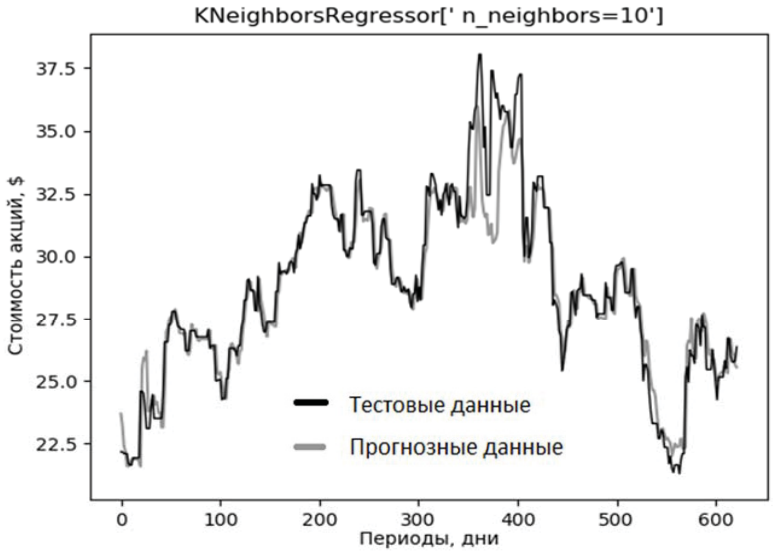

Рис. 5. Сравнение прогнозных и фактических цен акций Mail.ru Group методом стекинга на тестовой выборке

Fig. 5. Comparison of Forecast and Actual Mail.ru Group Share Prices by Stacking for the Test Sample

стекинговой ансамблевой модели, и фактические значения цены акций компании Mail.ru Group на тестовой выборке.

Применение ансамблей моделей может не превосходить по качеству прогноза отдельные алгоритмы. Результаты решения задачи прогнозирования для Mail.ru Group в случае применения индивидуального алгоритма SVM по некоторым оценкам (NRMSE, RAE, RRSE) более точные. В данном случае пользователю после рекомендаций о выборе лучшей модели предоставляется возможность дополнительного исследования параметров моделей и последующего тестирования.

Результаты

Применение ансамблей в процессе решения задачи прогнозирования позволяет повысить точность моделей по сравнению с индивидуальными алгоритмами. Использование подобных сложных моделей для нестационарных временных рядов может обеспечивать получение более качественного прогноза, что особенно важно в задачах анализа финансового рынка. Разработанное web-приложение позволяет построить разные ансамбли моделей с возможностью применения различных базовых методов. Приложение позволяет сравнить данные методы с применением расширенного набора функционалов ошибки. Оно упрощает и облегчает для пользователей некоторые этапы построения ансамблей моделей, предоставляя интерфейс для реализации вычислительных экспериментов.

Список литературы

1. *Чернышова Г. Ю.* Применение гибридных моделей для решения задач прогнозирования // Саратовской области – 80 лет : история, опыт развития, перспективы роста : сб. науч. тр. по итогам Междунар. науч.-практ. конф. : в 3 ч. Саратов : ССЭИ РЭУ им. Г. В. Плеханова, 2016. Ч. 2. С. 109–110.
2. *Zhang Ch., Ma Yu.* Ensemble Machine Learning. Methods and Applications. N. Y. : Springer Science Business Media, 2012. 331 p.
3. *Seni G., Elder J.* Ensemble Methods in Data Mining : Improving Accuracy Through Combining Predictions. San Rafael : Morgan & Claypool Publishers, 2010. 126 p.
4. *Zhou Z.* Ensemble Methods Foundations and Algorithms. Taylor & Francis Group, Chapman & Hall, 2012. 233 p.
5. *Shalev-Shwartz S., Ben-David S.* Understanding Machine Learning from Theory to Algorithms. Cambridge University Press, 2014. 449 p.
6. *Sagi O., Rokach L.* Ensemble learning : A survey. John Wiley and Sons, 2018. 18 p.
7. Yahoo Finance. URL: https://finance.yahoo.com/ (дата обращения: 15.10.2018).

Образец для цитирования:

Чернышова Г. Ю., Самаркина Е. А. Методы интеллектуального анализа данных для прогнозирования финансовых временных рядов // Изв. Сарат. ун-та. Нов. сер. Сер. Экономика. Управление. Право. 2019. Т. 19, вып. 2. С. 181–188. DOI: https://doi.org/10.18500/1994-2540-2019-19-2-181-188

Data Mining Methods for Financial Time Series Forecasting

G. Yu. Chernyshova, E. A. Samarkina

Galina Yu. Chernyshova, https://orcid.org/0000-0002-6464-0408, Saratov State University, 83 Astrakhanskaya St., Saratov 410012, Russia, galacherny@yandex.ru

Ekaterina A. Samarkina, https://orcid.org/0000-0001-7377-8444, Saratov State University, 83 Astrakhanskaya St., Saratov 410012, Russia, sea94@inbox.ru

Introduction. Data Mining algorithms enhancement leads to the solution of predictive analytics tasks in more efficient ways. The models ensembles are one of the actively developing areas, especially in those branches where predictive accuracy is more important than the interpretability of the model. **Theoretical analysis.** The forecasting problem requires careful study of the data set and methods suitable for analysis. It includes the solution of such subtasks as the choice of a forecasting model, the analysis of the forecast accuracy. Models ensembles are used to combine the predictions of several basic models in order to reduce forecasting errors and increase the generalizing ability of the individual models. Conceptual schemes are presented for the basic ensemble methods (bagging, boosting, stacking, blending). **Empirical analysis**. The practical aspects of forecasting financial time series, implementation of several models ensemble for forecasting problems are explored in the article. **Results**. To build models ensembles, we developed an application with a web interface that provides the ability to evaluate models with different error metrics, choose a more accurate models ensemble. With the web application, users can select and configure models to form forecasting ensembles, test the resulting models, and visualize the results. Testing of the models ensemble for the analysis of share prices time series is presented.

Keywords: forecasting, financial time series, Data Mining, models ensembles.

References

1. Chernyshova G. Yu. Primenenie gibridnykh modelei dlia resheniya zadach prognozirovaniya [Application of Hybrid Models for Forecasting Problems Solving]. In: *Saratovskoi oblasti – 80 let: istoriya, opyt razvitiya, perspektivy rosta. Trudy Mezhdunarodnoi nauchno-prakticheskoi konferentsii* [Saratov region – 80 years old: history, development experience, growth prospects. Proc. Int. sci. and pract. conf.: in 3 pt.]. Saratov, 2016, pt. 2, pp. 109–110 (in Russian).
2. Zhang Ch., Ma Yu. *Ensemble Machine Learning. Methods and Applications*. New York, Springer Science Business Media, 2012. 331 p.
3. Seni G., Elder J. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. San Rafael, Morgan & Claypool Publishers, 2010. 126 p.
4. Zhou Z. *Ensemble Methods Foundations and Algorithms*. Taylor & Francis Group, Chapman & Hall, 2012. 233 p.
5. Shalev-Shwartz S., Ben-David S. *Understanding Machine Learning from Theory to Algorithms*. Cambridge University Press, 2014. 449 p.
6. Sagi O., Rokach L. *Ensemble learning: A survey*. John Wiley and Sons, 2018. 18 p.
7. *Yahoo Finance*. Available at: https://finance.yahoo.com/ (accessed 18 December 2018).

Cite this article as:

Chernyshova G. Yu., Samarkina E. A. Data Mining Methods for Financial Time Series Forecasting. *Izv. Saratov Univ. (N. S.), Ser. Economics. Management. Law,* 2019, vol. 19, iss. 2, pp. 181–188 (in Russian). DOI: https://doi.org/10.18500/1994-2540-2019-19-2-181-188