

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра дискретной математики и информационных технологий

**ПРИМЕНЕНИЕ АНСАМБЛЕЙ МОДЕЛЕЙ
ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ДЛЯ ЗАДАЧИ
ПРОГНОЗИРОВАНИЯ**

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Самаркиной Екатерины Александровны

Научный руководитель

доцент каф. ДМиИТ, к. э. н. _____

Г. Ю. Чернышова

Заведующий кафедрой

доцент каф. ДМиИТ, к. ф.-м. н. _____

Л. Б. Тяпаев

Саратов 2019

ВВЕДЕНИЕ

Прогнозирование представляет собой оценку наиболее общих показателей перспективного развития. При решении предикативных задач 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 в этой области является получение разумно точных прогнозов будущих данных при анализе имеющейся информации. На сегодняшний день применение ансамблей моделей интеллектуального анализа данных является одной из актуальных проблем в прогностических задачах.

Отдельные методы интеллектуального анализа данных не всегда позволяют построить достаточно точные модели. К настоящему времени разработано множество различных методов и алгоритмов формирования ансамблей. Среди них наибольшее распространение получили такие методы, как бэггинг, бустинг, стекинг и блендинг.

Целью магистерской работы является программная реализация web-приложения для применения ансамблевых моделей интеллектуального анализа данных в задачах прогнозирования.

При этом необходимо решить следующие задачи:

- провести анализ алгоритмов построения ансамблевых моделей для прогнозирования;
- выбрать инструментарий для реализации ансамблей моделей;
- разработать web-приложение для формирования ансамблей моделей;
- выполнить вычислительный эксперимент для прогнозирования экономических показателей, оценить и сравнить результаты применения различных ансамблей моделей.

В разделе 1 представлен анализ современных ансамблевых моделей, а также методов прогнозирования. В разделе 2 описаны основные элементы web-приложения, реализующего ансамблевые модели для решения задач прогнозирования с использованием методов машинного обучения. В разделе 3 выполнен прогностический эксперимент, осуществлена оценка и сравнение

полученных результатов для выбора наиболее точной прогностической модели.

Методологические основы ансамблевых моделей представлены в работах Khairalla M. A., Ning X., AL-Jallad N. T., El-Faroug M. O., Neupane B., Woon W. L., Aung Z., Nooteboom El N. P. D., Feng Q. Y., Lopez C., Hernandez-Garcia E., Dijkstra H. A., Grmanova G., Laurinec P., Rozinajova V., Ezzeddine A. B., Lucka M., Lacko P., Vrablecova P., Navrat P., Divina F., Gilson A., Gomez-Vela F., Torres M. G., Torres J. F., Aydoqmus H. Y., Ekinici A., Erdal H., Erdal H., Foroozand H., Weijs St. V., Wichard J. D., Ogorzalek M [1, 8].

Практическая значимость магистерской работы заключается в программной реализации web-приложения для применения ансамблевых моделей интеллектуального анализа данных в задачах прогнозирования.

Структура и объём работы. Магистерская работа состоит из введения, 3 разделов, заключения, списка использованных источников и 1 приложения. Общий объем работы – 53 страницы, из них 52 страницы – основное содержание, включая 19 рисунков и 4 таблицы, цифровой носитель в качестве приложения, список использованных источников информации – 32 наименования.

1 КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Использование ансамблевых моделей в прогностических задачах» посвящен анализу ансамблевых моделей в прогностических задачах. На сегодняшний день применение ансамблей моделей интеллектуального анализа данных является одной из актуальных проблем в сфере информационных технологий, что не остается без внимания исследователей. Это очень активная исследовательская тема, которая охватывает различные области науки, техники, финансов.

Эффективной альтернативой использованию единственного метода прогнозирования является объединение прогнозов из нескольких разных моделей. Комбинация нескольких прогнозов почти всегда существенно снижает общие ошибки прогнозирования, а также превосходит модели компонентов.

Ансамблевые методы — это методы машинного обучения, которые объединяют несколько базовых моделей, чтобы создать одну оптимальную прогностическую модель. Система на основе ансамбля получается путем объединения различных моделей.

В последнее десятилетие ансамбли моделей стали областью очень активных исследований в машинном обучении, что привело к разработке большого числа разнообразных методов формирования ансамблей.

К настоящему времени разработано множество различных методов и алгоритмов формирования ансамблей. Среди них наибольшее распространение получили такие методы, как бэггинг, бустинг, стекинг и блендинг.

Первым вопросом при формировании ансамбля является выбор базовой модели (base model). Ансамбль в целом может рассматриваться как сложная, составная модель (multiple model), состоящая из отдельных (базовых) моделей. Здесь возможны два случая.

- ансамбль состоит из базовых моделей одного типа, например, только из деревьев решений, только из нейронных сетей и т.д.;
- ансамбль состоит из моделей различного типа – нейронных сетей, деревьев решений, регрессионных моделей и т.д.

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.

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. 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. Для агрегирования результатов базовых моделей в прогнозных задачах используется усреднение с использованием среднего арифметического значения или средневзвешенного значения.

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

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

Блендинг (blending) — простейшая схема стекинга: обучающую выборку делят на две части. На первой обучают базовые алгоритмы. Затем получают предсказания на основе них на второй части и на тестовой выборке. Предсказание каждого алгоритма можно рассматривать как новый признак (т.н. метапризнак). На метапризнаках второй части обучения настраивают метаалгоритм. Затем запускают его на метапризнаках теста и получают предсказание. Таким образом, можно сделать экспериментальный анализ следующим образом:

1. разбить обучающую выборку на два непересекающихся подмножества;

2. обучить несколько базовых моделей на первом подмножестве;
3. тестировать базовые модели на втором подмножестве;
4. обучить метаалгоритм, используя предсказания из предыдущего пункта как входные данные, а истинные классы объектов — как выход.

Блендинг и стекинг являются двумя очень похожими подходами. Часто в литературе эти термины часто используются как синонимы. Такие подходы обучают первый слой моделей и используют их выходы (т.е. вероятности) для обучения второго уровня моделей. Можно использовать любое количество слоев. Окончательное предсказание обычно выполняется средним правилом или конечным базовой моделью.

Выбор ансамблей модели — это сложная задача, т.к. их достаточно большое количество, они многопараметрические, оценка при прогнозировании осуществляется целым набором функционалов ошибки. Разработка инструментальных средств для реализации подбора ансамблей моделей и осуществление прогнозирования является актуальной задачей.

Благодаря интеллектуальному анализу данных можно получать предположения, которые основываются на большом количестве накопленных статистических данных. Интеллектуальный анализ данных (Data Mining) — исследование и обнаружение машиной (алгоритмами, средствами искусственного интеллекта) в данных скрытых знаний, которые: ранее не были известны, нетривиальны, практически полезны.

Эффективной альтернативой использованию единственного метода прогнозирования является объединение прогнозов из нескольких разных моделей. В связи с тем, что комбинация нескольких прогнозов почти всегда существенно снижает общие ошибки прогнозирования, а также превосходит модели компонентов.

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.

Основные этапами Data Mining являются:

- постановка задачи анализа;
- сбор данных для анализа;

- выбор методов интеллектуального анализа данных для построения моделей;
- оценка построенных моделей;
- интерпретация и использование моделей.

Методы интеллектуального анализа данных находят свое применение в бизнес-проектах, телекоммуникациях, производственном планировании, медицинских исследованиях.

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 случае категориальных значений — 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. В самых общих чертах решение задачи прогнозирования сводится к решению таких подзадач:

- выбор модели прогнозирования;
- анализ адекватности и точности построенного прогноза.

Рассмотрим комбинацию распространенных методов интеллектуального анализа данных, позволяющих решить задачу прогнозирования.

Метод ближайшего соседа (K-Nearest neighbour) относится к классу методов, работа которых основывается на хранении данных в памяти для сравнения с новыми элементами. При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется.

Дерево принятия решений (Decision tree) — это непараметрический метод обучения, используемый для классификации и регрессии. Цель состоит в том, чтобы создать модель, которая прогнозирует значение целевой переменной путем изучения простых правил принятия решений, выведенных из функций данных.

Стохастический градиентный спуск (Stochastic Gradient Descent, SGD) — это простой, но очень эффективный подход к дискриминационному обучению линейных классификаторов при использовании функций выпуклых потерь, таких как (линейные) машины опорных векторов и логистиче-

ская регрессия.

Метод опорных векторов (Support Vector Machine Regression, SVM) — это специфический класс алгоритмов, характеризующийся использованием ядер, отсутствием локальных минимумов, разреженности решения и т. д. Метод представляют собой набор контролируемых методов обучения, используемых для классификации, регрессии и обнаружения выбросов.

Для оценки моделей прогнозирования используются различные метрики качества, определяющие различие между полученным в результате применения модели значением и реальным значением. В разделе приведена таблица расширенного набора функций ошибки (Коэффициент детерминации (R^2), MSE, RMSE, NRMSE, RAE, RRSE, MAPE, SMAPE).

Отдельные функционалы оценки регрессионных моделей не дают полной картины, для решения практических задач целесообразно использовать несколько показателей. Выбор показателей для оценки прогностических моделей представляет собой отдельный этап предикативного анализа.

Второй раздел «Backend-разработка приложения для использования ансамблевых моделей» посвящен реализации web-приложения для использования ансамблевых моделей. Для разработки web-приложения был выбран язык программирования Python, т.к. он имеет широкий спектр библиотек, включающий в себя множество методов машинного обучения, такие как регрессия на основе k -ближайших соседей, регрессия на основе стохастического градиентного спуска, регрессия на основе деревьев решений, регрессия на основе метода опорных векторов.

В качестве методов прогнозирования использованы библиотеки sklearn. neighbors. KNeighborsRegressor (регрессия на основе k -ближайших соседей), sklearn. linear_model. SGDRegressor (регрессия на основе стохастического градиентного спуска), sklearn. tree. DecisionTreeRegressor (регрессия на основе деревьев решений), sklearn. svm. LinearSVR (регрессия на основе метода опорных векторов).

В настоящем приложении используются технологии стандартной трехуровневой архитектуры (CSS, HTML, JavaScript), а также библиотека JavaScript React. Приложение для построения ансамблевых моделей разработано с web-интерфейсом на языке программирования Python с использованием библиотек scikit-learn, Numpy, csv. Web-интерфейс приложения реализован на React.

Приложение дает возможность осуществить выбор модели для прогнозирования. Оно упрощает и облегчает для пользователей некоторые этапы построения ансамблей моделей при проведении эксперимента.

Разработанное web-приложение реализует следующие функциональные возможности:

1. загрузка файла с входными данными;
2. выбор ансамблей моделей;
3. задание параметров ансамбля;
4. выбор метода прогнозирования;
5. задание параметров для реализации методов прогнозирования;
6. формирование выборки;
7. визуализация результатов прогнозирования на основе построенной модели;
8. прогнозирование на основе выбранной модели.

Приложение позволяет построить ансамбли моделей с выбранными методами. Увидеть функционалы качеств моделей и выбрать наиболее эффективную модель.

Третий раздел «Применение ансамблевых моделей для прогнозирования фондового рынка» посвящен применению ансамблевых моделей для прогнозирования фондового рынка. Прогноз фондового рынка — это попытка определить будущую стоимость акций компании или другого финансового инструмента, торгуемого на финансовой бирже. Методы машинного обучения могут раскрыть модели и идеи, которые можно использовать для безошибочно точных прогнозов. В целом анализ фондового рынка делится на две части — фундаментальный анализ и технический анализ. При прогнозировании с помощью алгоритмов машинного обучения существует возможность вычислить будущую цену акций. Как только алгоритм разработан, его можно использовать на тестовых данных и анализировать производительность алгоритма, сравнивая прогнозируемую тенденцию с фактической производительностью акций.

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

основе минимизации функционалов ошибки. 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 большому 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. 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, дерева 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. Для визуализации результатов прогнозирования приложение позволяет построить графики, обеспечивающие возможность пользователю в процессе оценки модели получить дополнительную информацию о качестве построенной модели.

Наилучшей моделью будет являться модель с наименьшими функционалами качества, которую можно определить методом сравнения и анализа всех функционалов качества для построенных ансамблей моделей.

ЗАКЛЮЧЕНИЕ

Применение ансамблей в процессе решения задачи прогнозирования в ряде случаев 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 различные ансамбли моделей с возможностью применения 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.

В работе был произведен анализ алгоритмов построения ансамблевых моделей для прогнозирования. Созданное web-приложение, где были собраны основные методы построения ансамблей моделей, позволяет определить наиболее эффективную модель для решения прогностической задачи.

В качестве инструментария для реализации web-приложения для формирования ансамблей моделей был выбран язык программирования Python для рассматриваемых ансамблей моделей, библиотека JavaScript React для создания web-интерфейса. Python обеспечивает комплексное решение, в нем реализован широкий спектр библиотек, включающий в себя множество методов машинного обучения, такие как регрессия на основе k -ближайших соседей, регрессия на основе стохастического градиентного спуска, регрессия на основе деревьев решений, регрессия на основе метода опорных векторов, а также его можно подключить к web-интерфейсу.

При использовании созданного web-приложения был выполнен вычислительный эксперимент для прогнозирования экономических показателей, а также для оценки фондового рынка. Было реализовано сравнение результатов применения различных ансамблей моделей.

В результате использования разработанного приложения построения для ансамблей моделей достигается возможность получения наиболее точной модели для прогнозирования.

По материалам работы была опубликована статья Чернышова Г. Ю., Самаркина Е. А. Методы интеллектуального анализа данных для прогнози-

рования финансовых временных рядов // Известия Саратовского университета, Новая серия. Серия Экономика. Управление. Право, Научный журнал: ФГБОУ ВО СГУ им. Н. Г. Чернышевского, 2019. Выпуск № 2. Том 19, – 181-189 с. Поставленные цель и задачи магистерской работы были выполнены полностью.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 Khairalla M. A., Ning X., AL-Jallad N. T., El-Faroug M. O. Short-Term Forecasting for Energy Consumption through Stacking Heterogeneous Ensemble Learning Model // M. A. Khairalla, X. Ning, N. T. AL-Jallad, M. O. El-Faroug / MDPI, Basel, Switzerland; 2018. – 21 p.
- 2 Neupane B., Woon W. L., Aung Z. Ensemble Prediction Model with Expert Selection for Electricity Price Forecasting // B. Neupane, W. L. Woon, Z. Aung / MDPI, Basel, Switzerland; 2017. – 28 p.
- 3 Nooteboom El N. P. D., Feng Q. Y., Lopez C., Hernandez-Garcia E., Dijkstra H. A. Using Network Theory and Machine Learning to predict // El N. P. D. Nooteboom, Q. Y. Feng, C. Lopez, E. Hernandez-Garcia, H. A. Dijkstra / Earth System Dynamics 9, 969-983; 2018. – 15 p.
- 4 Grmanova G., Laurinec P., Rozinajova V., Ezzeddine A. B., Lucka M., Lacko P., Vrabecova P., Navrat P. Incremental Ensemble Learning for Electricity Load Forecasting // G. Grmanova, P. Laurinec, V. Rozinajova, A. B. Ezzeddine, M. Luck, P. Lacko, P. Vrabecov, P. Nvrat / Acta Polytechnica Hungarica Vol. 13, No. 2; 2016. – 21 p.
- 5 Divina F., Gilson A., Gomez-Vela F., Torres M. G., Torres J. F. Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting // F. Divina, A. Gilson, F. Gomez-Vela, M. G. Torres, J. F. Torres / Licensee MDPI, Basel, Switzerland; 2018. – 31 p.
- 6 Aydoqmus H. Y., Ekinici A., Erdal H., Erdal H. Optimizing the monthly crude oil price forecasting accuracy via bagging ensemble models // H. Y. Aydoqmus, A. Ekinici, H. I. Erdal, H. Erdal / Journal of Economics and International Finance; 2015. – 127-136 p.
- 7 Foroozand H., Weijs St. V. Entropy Ensemble Filter: A Modified Bootstrap Aggregating (Bagging) Procedure to Improve Efficiency in Ensemble Model Simulation // H. Foroozand and St. V. Weijs / Department of Civil Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; 2017. – 16 p.
- 8 Wichard J. D., Ogorzalek M. Time Series Prediction with Ensemble Models // Jorg D. Wichard, Maciej Ogorzaek / Conference: Neural Networks.

Proceedings. 2004 IEEE International Joint Conference on, Volume: 2. 2004.
– 6 p.