МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования   
**«Нижегородский государственный университет им. Н.И. Лобачевского»**

**Национальный исследовательский университет**

#### **Институт информационных технологий, математики и механики**

#### **Кафедра: Математического обеспечения**

#### **и суперкомпьютерных технологий**

Направление подготовки: «Фундаментальная информатика

и информационные технологии»

**МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ**

Тема:

**«Инструментальная СИАД. Разработка библиотеки нейросетей»**

Заведующий кафедрой: Выполнил:

д.ф.-м.н., проф., студент группы 8403

Стронгин Роман Григорьевич Смирнов Михаил Александрович

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ подпись подпись

Научный руководитель:

к.т.н., доц.,

Карпенко Сергей Николаевич

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

Нижний Новгород  
20\_\_

Оглавление

[Введение 3](#_Toc438055573)

[Список литературы 5](#_Toc438055574)

# Введение

Повсеместное использование электронных вычислительных средств во многих сферах деятельности человека привело к стремительному росту объемов хранимой и обрабатываемой информации. Общество генерирует гигабайты текстовых файлов, видео, фото и аудио-контента ежедневно благодаря тесной интеграции всемирной сети «Интернет» в его жизнь. Однако большая часть производимой обществом информации находится в «сыром», не стандартизованном, сложном для анализа виде. Такую информацию принято называть **данными** [1]. Их анализ представляет собой довольно трудоемкую задачу: данные разнородны, содержат большое количество внутренних зависимостей, ошибок и**шума** (информации, не влияющей на результат анализа).

**Интеллектуальный анализ данных** (англ. Data Mining) занимается извлечением неявной, ранее неизвестной и потенциально полезной информации из больших наборов данных [1, 2]. Для этого применяются различные алгоритмы **искусственного интеллекта** [3] и **машинного обучения** [4]: нейронные сети [5], деревья решений [6], метод опорных векторов, случайный лес [7], байесовские сети и т.д. Эти методы хорошо зарекомендовали себя при решении трудно формализуемых задач [3, 8-10], к каким относится и задача выявления полезной информации в любой области знаний. Интеллектуальный анализ данных успешно применяется, например, при диагностике и предсказании сердечно-сосудистых заболеваний [11, 12], увеличении продаж на рынке мобильных телекоммуникаций [13], предсказании мест возникновения аварий [14], оценки индивидуальных лекарственных доз [15], анализе следственных версий [16].

Решению задачи анализа данных предшествуют стадии сбора и предобработки данных, выбора модели решения задачи и подбора ее параметров – **обучения** модели [4]. Большинство современных авторов работ в области интеллектуального анализа данных исследуют возможность использования того или иного метода решения задачи применительно к имеющимся данным; основой их работ является описание совершённых над исходными данными преобразований для улучшения предсказательных способностей решающей модели. Однако данные преобразования являются, в большинстве своем, эффективными только в рассматриваемой автором предметной области. Правда, существуют исследования, выявляющие правила предобработки данных, способствующие повышению эффективности моделей [17, 18]. Но попыток формулирования фундаментальных правил выбора наилучшей модели для решения задач конкретного класса не наблюдается. Большинство работ описывают достижения в решении задач классификации, кластеризации и восстановления зависимости. Некоторые работы предлагают эффективные решения задач анализа временных последовательностей [17]. Каждая из этих работ предлагает способы предобработки данных и обучения конкретной модели; но не делается попытки сравнения и анализа эффективности использования различных методов для решения конкретной задачи.

Эффективность применения методов интеллектуального анализа данных в самых разных научных и прикладных областях способствовало созданию множества различных программных средств и инструментов, позволяющих упростить проведение экспериментов и решение задач [19, 20]. Были сформулированы новые техники обработки больших массивов данных (например, использование распределенной структуры проведения вычислений в облачных серверах для обработки больших объемов данных [21]), новые методы обучения моделей и решения задач (глубокое обучение [22, 23]). Однако современные программные средства интеллектуального анализа данных нацелены на получение максимального результата для выбранного пользователем метода решения задачи.

Цель данной работы заключается в создании библиотеки нейронных сетей в рамках разработки инструментальной системы интеллектуального анализа данных (ИСИАД). Данная система разработана в команде (в которую также входит автор данной работы). ИСИАД включает в себя несколько различных методов решения задач (библиотека нейронных сетей, деревья решений) и алгоритмов обучения. Эта система автоматизирует процесс выбора наилучшей модели для решения задач **классификации** и **восстановления зависимости** [24]. Таким образом, актуальность данной работы состоит, во-первых, в упрощении использования методов интеллектуального анализа данных для решения поставленных задач: пользователь данной системы не обязан быть экспертом в области интеллектуального анализа данных, т.к. он не должен выбирать метод решения задачи вручную. Во-вторых, система актуальна с исследовательской стороны, т.к. для решения одной задачи будут применяться различные модели, методы обучения задачи и алгоритмы предобработки данных, будут доступны для анализа их параметры и мера эффективности каждой обученной модели. Автор данной работы надеется, что это поспособствует исследованию сравнительной эффективности методов при решении целых классов задач.

# Список литературы

1. Witten, I.H. Data Mining: Practical machine learning tools and techniques / I.H. Witten, E. Frank, M.A. Hall – Burlington: Elsevier, 2011.
2. Frawley, W.J. Knowledge Discovery in Databases: An Overview / W.J. Frawley, G. Piatetsky-Shapiro, C.J. Matheus // AI Magazine – 1992. – 13(3).
3. Russel, S.J. Artificial Intelligence: A Modern Approach / S.J. Russel, P. Norvig – 2nd edition – New Jersey: Prentice Hall, 2003.
4. Mitchell, T. Machine Learning // McGraw-Hill Science/Engineering/Math – 1997.
5. Haykin, S. Neural Networks. A comprehensive foundation / 2nd edition – New Jersey: Prentice Hall, 1999.
6. Breiman, L. Classification and regression trees / L. Breiman, J.H. Friedman, R.A. Olshen // Wadsworth & Brooks/Cole Advanced Books & Software – 1984.
7. Breiman, L. Random forests // Machine Learning – 2001. – 45.
8. Rosenblatt, F. The Perceptron: A probabilistic model for information storage and organization in the brain // Psych. Rev. – 1958. – 65.
9. Minsky, M.L. Perceptrons / M.L. Minsky, S.A. Papert // MIT Press – 1969.
10. Kohonen, T. Self-Organizing Maps // Springer – 1995.
11. Ansari, U. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction / U. Ansari, S. Soni // International Journal of Computer Applications – 2011.
12. Srinivas, K. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks / K. Srinivas, B. Kavihta Rani, A. Govrdhan // International Journal on Computer Science and Engineering – 2010. – 2.
13. Ahn, H. Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques / H. Ahn, J.J. Ahn, K.J. Oh, D.H. Kim // Expert Systems with Applications – 2011. – 38(5).
14. Rivas, T. Explaining and predicting workplace accidents using data-mining techniques / T. Rivas, M. Paz, J.E. Martin, J.M. Matias // Reliability Engineering and System Safety – 2011. – 96(7).
15. Ефимов, А.С. Метод оценки индивидуальных лекарственных доз на основе искусственной нейронной сети // Н.Новгород: изд. Нижегородского госуниверситета, 2011.
16. Изутов, Е.О. Анализ следственных версий на основе искусственной нейронной сети // Н.Новгород: изд. Нижегородского госуниверситета, 2013.
17. Fu, T.C. A review on time series data mining // Engeenering applications of artificial intelligence – 2011. – 24.
18. Liu, H. Feature Selection: An ever evolving frontier in Data Mining / H. Liu, H. Motoda, R. Setiono, Z. Zhao // JMLR: Workshop and Conference Proceedings – 2010. – 10.
19. Alcala-fdez, J. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework / J. Alcala-fdez, A. Fernandez, J.Luengo, J.Derrac, S.Garcia, L. Sanchez, F.Herrera // Old City Publishing, Inc. – 2011. – 17.
20. Cortez, P. Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool // University of Minho – 2010.
21. Low, Y. Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud // Carnegie Mellon University – 2012.
22. Dahl, G.E. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition // IEEE transactions on audio, speech, and language processing – 2012.
23. Krizhevsky, A. ImageNet classiﬁcation with deep convolutional neural networks // A. Krizhevsky, I. Sutskever, G.E. Hinton // University of Toronto – 2013.
24. Чубукова, И. А. Data Mining: учебное пособие / И. А. Чубукова. – М. : Интернет-университет информационных технологий; Бином. Лаборатория знаний, 2006.