Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМ. Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра компьютерной физики
и метаматериалов на базе Саратовского филиала
Института радиотехники и электроники
им. В.А. Котельникова РАН

ПОСТРОЕНИЕ И ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ МУЛЬТИНОМИАЛЬНОГО НАИВНОГО БАЙЕСОВСКОГО КЛАССИФИКАТОРА НАУЧНЫХ СТАТЕЙ

АВТОРЕФЕРАТ

ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ (МАГИСТЕРСКОЙ) РАБОТЫ

студента 2 курса 256 группы
направления 03.04.02 «Физика» физического факультета
Семендяева Даниила Алексеевича

Заведующий кафедрой
д. ф.-м. н., профессор _____В.М.Аникин

«08» 06.2020

Научный руководитель
д. ф.-м. н., профессор _____А.С.Ремизов

«08» 06.2020

Саратов
2020

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуализация работы. На современном этапе состояния общества информационные технологии (ИТ) породили все увеличивающийся поток разнородной информации. Основной задачей поисковых систем (поисковых машин) является предоставление качественных результатов, т.е. наиболее важных релевантных страниц. Для этого необходимо решать задачу классификации (classification problem). Поэтому теория, методы и алгоритмы классификации информации являются **актуальным** и бурно развивающимся научным направлением.

Классификация информации в сетях, и в частности в сети Интернет, позволяет решать различные задачи, например: документооборот, автоматическое аннотирование и реферирование, машинный перевод, составление интернет-каталогов, ограничение области поиска в поисковых системах, определение кодировки и языка текста, классификация новостей.

Машинное обучение концентрируется на разработке таких компьютерных программ и алгоритмов, которые сами учатся расти и адаптироваться при подаче новых данных. Этот процесс не похож на процесс интеллектуального анализа данных. Обе системы проходят через предоставленные им данные или собираются в поисках шаблонов. Однако в приложениях для интеллектуального анализа данных, данные извлекаются для понимания человеком, в то время как алгоритмы машинного обучения используют эти данные для поиска шаблонов в данных и соответственно изменения действий программы.

В машинном обучении классификацию понимают, как задачу определения класса для ранее не встречавшегося образца(объекта) на основе эмпирических данных, так называемых прецедентов, которые описывают исследуемые образцы и отражают присущие им свойства и закономерности. Существует зависимость между образцами и классами, но она неизвестна. Множество прецедентов, пар образец-класс, составляет обучающую выборку, по которой нахо-

дится зависимость, то есть строится алгоритм, способный для любого образца выдать ответ, к какому классу тот принадлежит. Это пример обучения с учителем. Под учителем в данном случае понимается обучающая выборка. Примерами таких моделей, основанных на машинном обучении, являются байесовские классификаторы. В работе рассмотрены различные модификации наивного байесовский классификатора, реализован мультиноминальный классификатор. В байесовских классификаторах используется критерий, минимизирующий вероятность принятия ошибочного решения, поэтому байесовские алгоритмы являются статистически оптимальными. Однако для этого алгоритмы требуют в идеале полного знания многомерных функций распределения наблюдаемых признаков для каждого класса. Необходимость такого знания обусловлена использованием формулы Байеса, которая лежит в основе байесовских методов принятия решения.

Целью работы является анализ эффективности мультиноминального наивного байесовского классификатора в задаче классификации научных статей. В задачи работы входит:

1) методическое изложение теоретической базы, необходимой в работе, рассмотрение различных моделей текста, вариантов байесовских алгоритмов классификации и методов оценки качества классификаторов;

2) поиск и предобработка наборов данных, содержащих тексты статей и из аннотации, программная реализация мультиномиального классификатора

3) валидация моделей, оценка качества классификации различных вариантов моделей, обученных на наборах данных разного объема, отдельно для обучения по аннотациям и по полным текстам статей, оценки времени обучения и времени классификации моделей.

СТРУКТУРА И ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении сформулированы аспектные характеристики работы - актуальность, цель ВКР, решаемые задачи.

В первой главе описаны различные математические модели текста, используемые в задачах компьютерной обработки информации.

Во второй главе представлен байесовский подход к классификации, основанный на теореме, утверждающей, что если плотности распределения каждого из классов известны, то искомый алгоритм можно выписать в явном аналитическом виде. Приведена теорема Байеса, позволяющая определить вероятность какого-либо события при условии, что произошло другое статистически взаимосвязанное с ним событие, часто используемая для обновления вероятности гипотезы по мере получения дополнительных данных.

Отражена задача классификации заключающаяся в том, чтобы построить алгоритм $\alpha: X \to Y$, способный классифицировать произвольный объект $x \in X$.

Рассмотрено, применение наивного байесовского классификатора в задаче классификации научных статей и его разновидности.

В третьей главе была выполнена программная реализация классификатора на языке python с использованием библиотек scikit-learn (Machine Learning in Python), numpy (для работы с матрицами) и pymongo (для работы с нереляционной базой данных mongodb), в среде разработки PyCharm.

Были определены числовые метрики, позволяющие оценить качество не зависимо от распределения примеров тестового множества, среди которых точность и полнота, матрица неточностей, F_1-мера.

Для выполнения работы были взяты примеры текстов аннотаций и текстов статей по тематикам «физика» и «медицина». По этим данным были обучены 12 вариантов моделей – 6 для текстов статей и 6 для текстов аннотаций. Объемы сбалансированных выборок, используемых для обучения – 2, 20, 200, 2000, 20000, 100000 статей и аннотаций соответственно. Валидация была проведена на сбалансированной выборке размером 100000 в двух вариантах – со смещением и без, для получения смещенной и несмещенной оценок относи-

тельно обучающего набора. Все результаты приведены для смещенной выборки. Добавлены расчеты матрицы ошибок, параметров P, R и f1 меры.

Результаты представлены на рисунках 1–5.

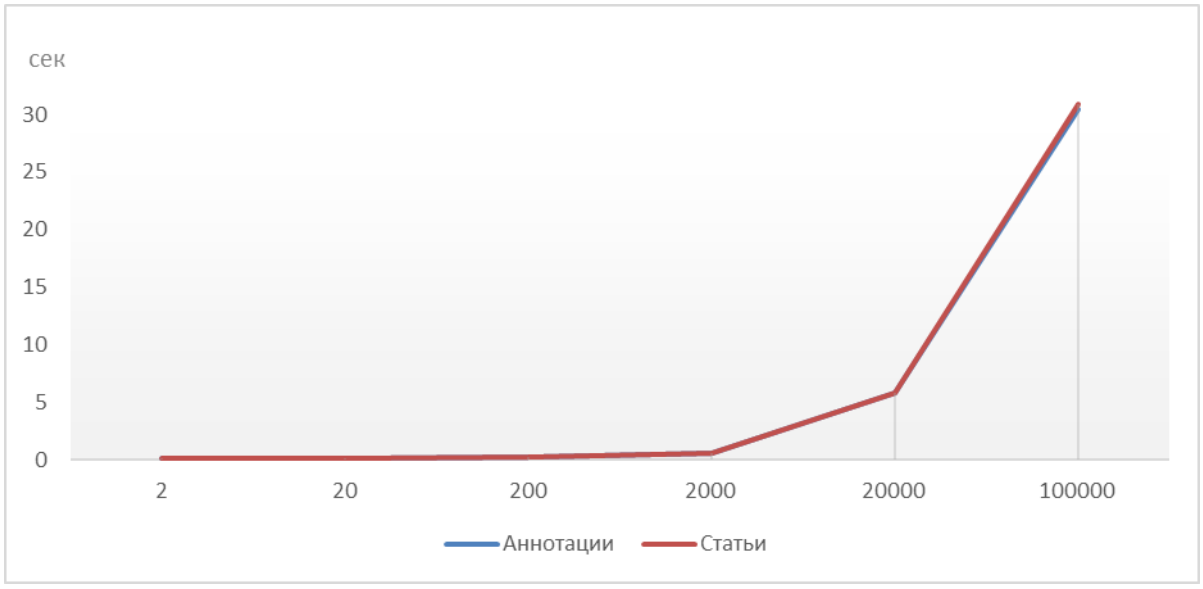

Рисунок 1 – Зависимость времени обучения от объема обучающей выборки

По графику видно, что время обучения моделей начинает значительно увеличиваться, когда количество документов, приближается к 2000. Для аннотаций и статей время обучения примерно одинаковое.

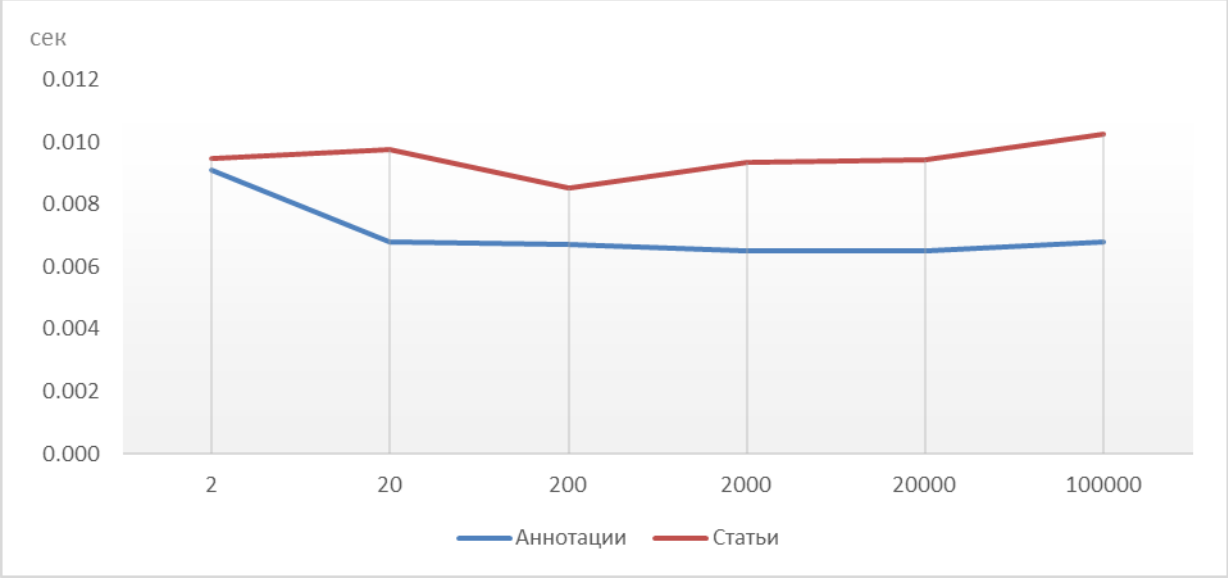

Рисунок 2 – Зависимость времени классификации одного документа (в секундах) от объема обучающей выборки

Исходя из графика видно, что наивысшая скорость классификации у моделей, обученных на 20000 (для аннотаций) и 200 (для статей) документов.

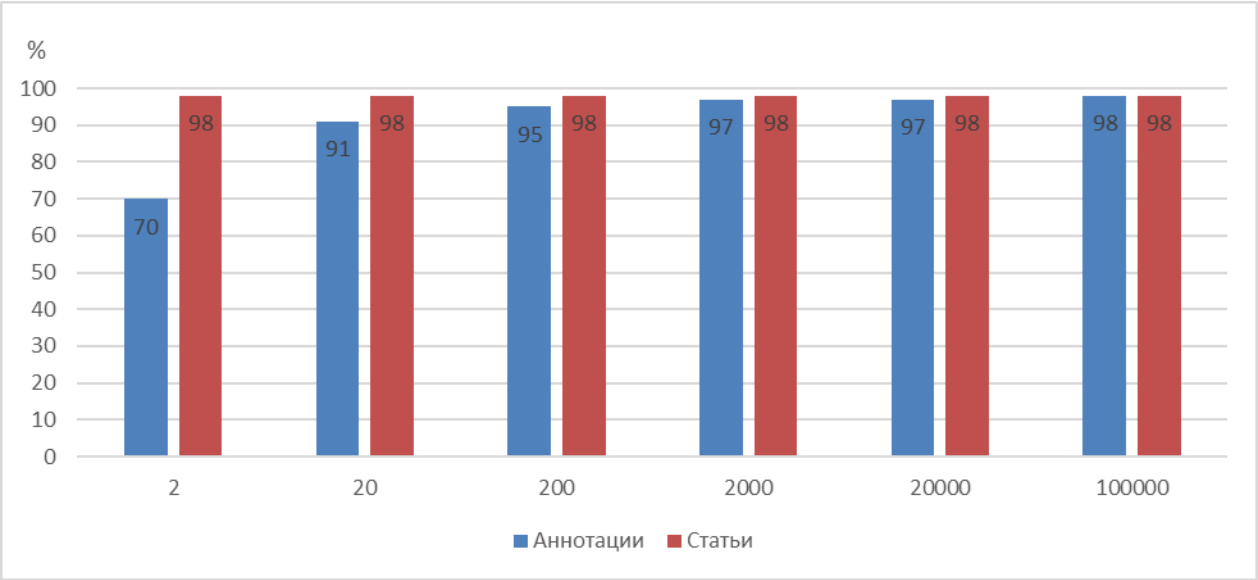

Рисунок 3 – Истинно положительная оценка по смещенной выборке

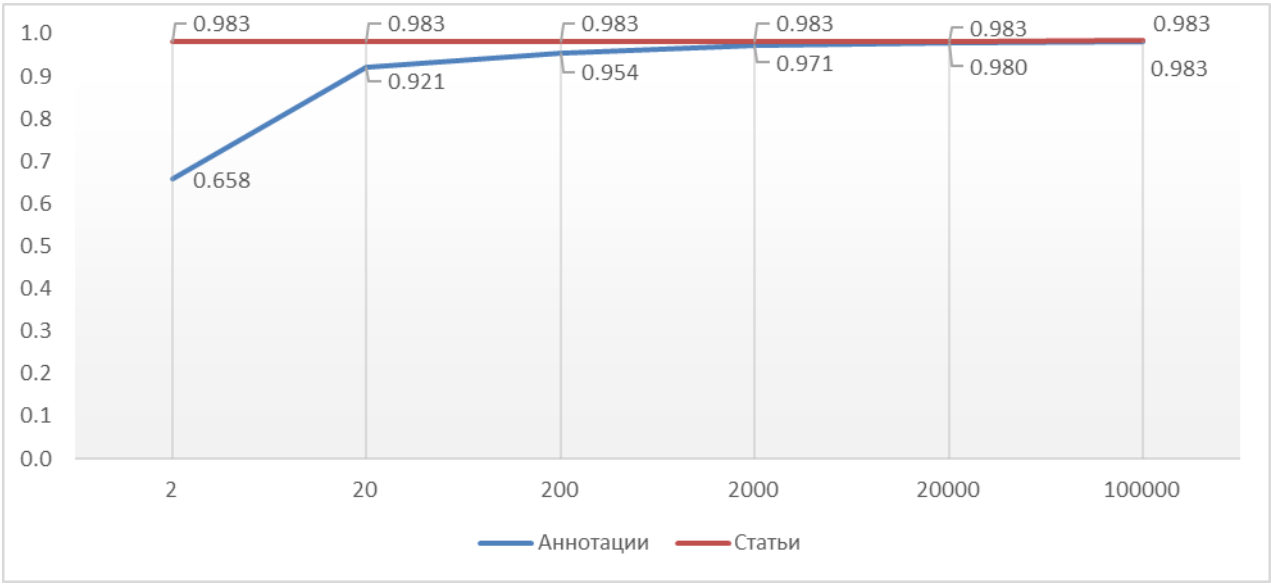

Рисунок 4 – F1 мера для данных «физика» по смещенной выборке

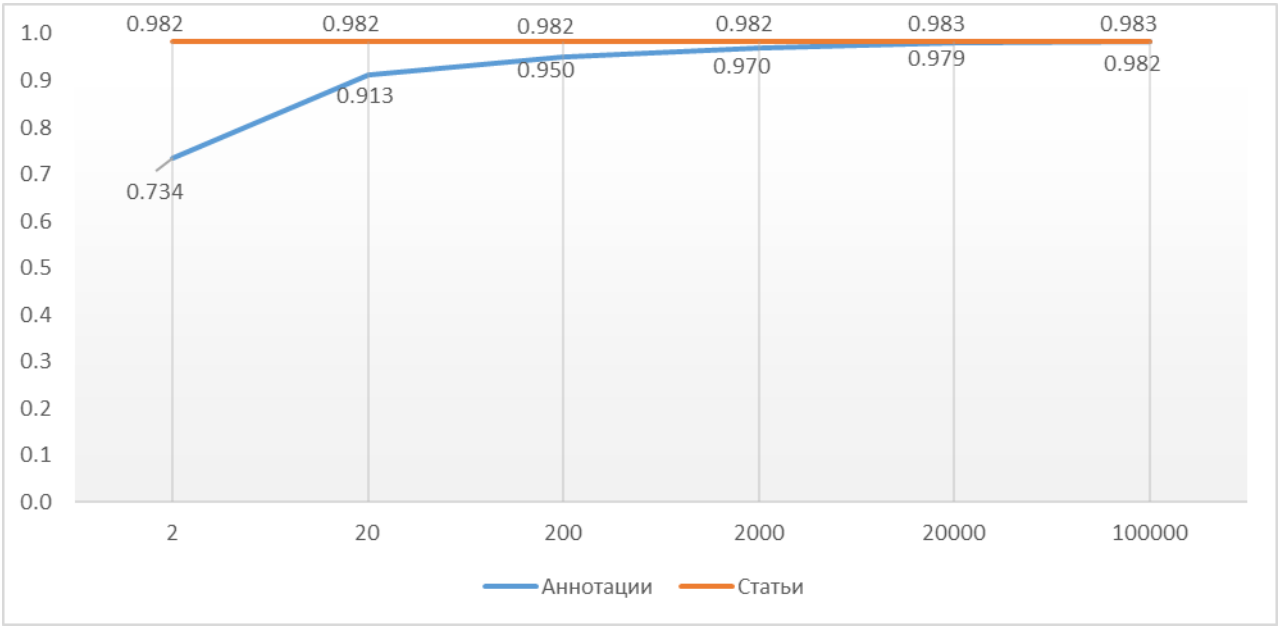

Рисунок 5 – F1 мера для данных «медицина» по смещенной выборке

Из данных графиков видно, что по f1-мере качество классификаторов выравнивается, начиная с объемов обучающей выборки от 20 тыс. статей. При объеме в 2 тыс. разница в одной десятой и может быть приемлемой. Однако при более меньших объемах вариант модели, обученный на полнотекстовых статьях, а не только на аннотациях, существенно выигрывает в качестве. При этом время обучения, согласно рис. 4, практически не отличается, но время оценивания одного документа несколько выше для варианта с аннотациями (7 миллисекунд против 10 для статей).

В **Заключении ВКР** сформулирован основные выводы по работе. е

ОСНОВНЫЕ ВЫВОДЫ

В теоретической части работы рассмотрены алгоритмы байесовской классификации и математические модели текста.

В практической части программно реализован мультиноминальный наивный байесовский классификатор, найден и адаптирован для классификации

набор данных, содержащий 200 тыс. научных статей по физике и медицие. Обучено 12 вариантов модели классификатора, различающиеся типом исходных данных (аннотации или сами статьи) и объемами обучающих выборок.

Проведена оценка качества классификации различных вариантов моделей, оценки времени обучения и времени классификации моделей. Валидация была проведена на отдельной (смещенной) выборке по 50000 документов на класс (всего 100тыс. статей), отличной от той, по которой проводилось обучение (также 100тыс.).

Для оценки качества моделей были вычислены следующие показатели: матрица ошибок, параметры P (точность), R (полнота) и f1 меры для каждого класса.

По результатам исследования можно сделать следующие **выводы**:

- при увеличении объема обучающей выборки время на обучение растет нелинейно, и при объемах до 2тыс. составляет около 200 мс, а при дальнейшем увеличении начинается экспоненциальный рост (до 30 секунд при максимальном объеме в 100 тыс.). При этом для аннотаций и статей время обучения примерно одинаковое

- среднее время классификации одной статьи составляет 7 мс для аннотаций и 9 мс для полных текстов статей, что можно считать сопоставимыми величинами;

- качество классификаторов выравнивается по f1-мере, начиная с объемов обучающей выборки от 20 тыс. статей. При объеме в 2 тыс. разница между классификаторами составляет около одной десятой и может быть приемлемой. При меньших объемах вариант модели, обученный только на аннотациях, существенно проигрывает в качестве.

С практической точки зрения, оптимальным выглядит вариант модели, обученной на полных текстах статей. В этом случае не требуется большой объем обучающей выборки, главное, чтобы она была сбалансирована относительно

классов, а скорость работы и обучения остается примерно такой же, как в более простом случае аннотаций.

Список использованных источников

1. Коршунов Антон, Гомзин Андрей. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. Т: 23/ С. 215 – 244

2. Jones K. S. A statistical interpretation of term specificity and its application in retrieval (англ.) // Journal of Documentation: журнал. — MCB University: MCB University Press, 2004. Vol. 60, no. 5. P. 493-502.

3. Солтон Дж. Динамические библиотечно-поисковые системы. М.: Мир, 1979.

4. Salton, G. and McGill, M. J. 1983 Introduction to modern information retrieval. McGraw-Hill, 1983.

5. Proceedings of the 7th Annual Conference ZNALOSTI 2008, Bratislava, Slovakia, February 2008. Pp. 54 – 65.

6. Jurafsky, D. and Martin, J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. — Pearson Prentice Hall, 2009. — 988 p. — ISBN 9780131873216.

7. Proceedings of the ITAT 2008, Information Technologies — Applications and Theory, Hrebienok, Slovakia, , September 2008. 2008. Pp. 23-26.

8. Tomas Mikolov, Quoc Le. Distributed Representations of Sentences and Documents. // Proceedings of Workshop at The 31st International Conference on Machine Learning (ICML). 2014.

URL : http://jmlr.org/proceedings/papers/v32/le14.pdf

9. Dietterich T.G. Machine learning research: four current directions // AI Magazine, 1997. V. 18. P. 97-136

10. Quinlan J.R. Bagging, Boosting, and C4.5 // Proceedings of AAA/IAAI. 1996. P. 725-730.

11. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. М.: Финансы и статистика, 1989.

12. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974.

13. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979.

14. Дуда Р., Харт П. Распознавание образов и анализ сцен. М.: Мир, 1976.

15. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2001.

16. Zhang H. The optimality of Naive Bayes // Proc. FLAIRS. 2004.

17. Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers // ICML 2003. Vol. 3. Pp. 616-623.

18. C.D. Manning, P. Raghavan and H. Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008. P. 234-265.

19. McCallum A. and Nigam K. A comparison of event models for Naive Bayes text classification. Proc. AAAI/ICML-98. 1998. Workshop on Learning for Text Categorization. Pp. 41-48.

20. Metsis V., Androutsopoulos I. and Paliouras G.. Spam filtering with Naive Bayes – Which Naive Bayes? 3rd Conf. on Email and Anti-Spam (CEAS). 2006.

21. Кобзарь А. И. Прикладная математическая статистика. М.: Физматлит, 2006. 816 с.

22. Крамер Г. Математические методы статистики. — М.: Гос. изд-во иностр. лит-ры, 1948. 631 с.

23.. Справочник по теории вероятностей и математической статистике / под ред. В.С. Королюка. Киев: Наукова думка, 1978. 582 с.

24. Жуков Д. А. Анализ критериев качества классификации при диагностике функционирования технического объекта // Математическое моделирование. 2019, № 3(57). С. 112 – 117.