Министерство науки и высшего образования РФ

Федеральное государственное автономное
образовательное учреждение высшего образования

«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Кафедра высшей и прикладной математики

УТВЕРЖДАЮ

Заведующий кафедрой

_____/ С.Г. Мысливец

«____» _____ 2023 г.

БАКАЛАВРСКАЯ РАБОТА

Направление 01.03.02 Прикладная математика и информатика

МЕТОДЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ МАЛЫХ ВЫБОРОК ДАННЫХ

Руководитель _____ доцент, кандидат физико-математических наук Д.В. Семенова

Выпускник _____ Д.И. Ильина

Нормоконтролер _____ Т.Н. Шипина

Красноярск 2023

РЕФЕРАТ

Выпускная квалификационная работа по теме "Методы статистической обработки малых выборок данных" содержит 60 страниц текста, 12 использованных источников.

МАЛАЯ ВЫБОРКА, ОЦЕНКА ФУНКЦИИ ПЛОТНОСТИ, СТАТИСТИЧЕСКИЙ АНАЛИЗ МАЛЫХ ВЫБОРОК.

Цель работы — исследование методов восстановления функции плотности распределения вероятности посредством методов для работы с малыми выборками данных.

В результате исследований проведен обзор методов восстановления функции плотности по выборке, реализованы методы Розенблатта–Парзена, проекционной оценки, Карандеева–Эйсымонта и метод гистограмм, проведен сравнительный анализ методов восстановления функции плотности распределения вероятности на модельных данных для малых выборок, выполнены вычислительные эксперименты с целью восстановления функции плотности распределения вероятностей на модельных данных стандартного нормального и экспоненциального распределений, бета-распределения и распределения Парето. А также проведены вычислительные эксперименты на реальных данных, где в качестве случайной величины использовался ключевой показатель эффективности сотрудника (KPI).

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

На практике довольно часто приходится иметь дело с выборками весьма малого объема, численности которых меньше двадцати-тридцати. Но на самом деле выборку следует считать малой, если при ее обработке методами, основанными на группировке данных, нельзя достичь заданных точности и достоверности [8]. И для анализа выборок такого объема мы не можем использовать классическую математическую статистику, поскольку полученные результаты будут попросту недостоверны.

Целью данной работы является поиск и реализация методов, которые подходят для обработки малых выборок данных. С помощью этих методов решается задача восстановления плотности распределения вероятностей на основе малых выборок данных. А также проводится сравнение реализации методов на модельных и реальных данных.

В работе исследованы и программно реализованы следующие методы:

- метод проекционной оценки;
- метод гистограмм;
- метод Карандеева-Эйсымонта;
- метод Розенблатта–Парзена.

Вычислительные эксперименты проводились на данных показателей эффективности торговых представителей Сибирской кондитерской компании. По анализу всех критериев оценивания торговых представителей был рассчитан ключевой показатель эффективности каждого сотрудника, который в дальнейшем использовался для восстановления плотности распределения вероятностей. Полученную функцию плотности распределения вероятностей можно исследовать в задаче прогнозирования, например, с какой вероятностью ключевой показатель эффективности конкретного сотрудника попадет в тот или иной интервал.

ЗАКЛЮЧЕНИЕ

В первой главе были рассмотрены все методы, используемые в данной работе, и некоторые дополнительные методы, изученные в ходе работы. Все эксперименты во второй и третьей главах проводились только на некоторых методах: метод Розенблатта–Парзена, метод гистограмм, метод проекционной оценки и метод Карандеева–Эйсымонта. Также в таблице 1.4 приведены достоинства и недостатки каждого из реализованных методов.

В второй главе были рассмотрены вычислительные эксперименты на модельных данных на бета, экспоненциальном, стандартном нормальном и Парето распределениях. Приведены все графики по каждому из реализованных методов, а также посчитана среднеквадратичная ошибка для различных объемов выборки в таблице 2.1.

Из полученных результатов в таблице 2.1 можно сделать вывод о том, что метод гистограмм и метод Розенблатта–Парзена практически не чувствительны к объему исследуемой выборки и показывают почти точный результат. Методы прекционной оценки и Карандеева–Эйсымонта наоборот показывают результат лучше, когда исследуемая выборка имеет относительно малый объем.

В третьей главе было введено понятие KPI и проведены вычислительные эксперименты на реальных данных от Сибирской кондитерской компании. Были получены значения KPI для трех сотрудников компании помесячно и исходя из полученных результатов, можно наблюдать насколько хорошо работал сотрудник в определенный месяц года.

Также было проведено сравение по плотности распределения KPI для трех сотрудников. Из данных этих графиков можно увидеть, что чем уже плотность распределения (то есть, чем меньше у нее разброс), тем стабильнее работает сотрудник.

Итак, поставленная цель выполнена: найдены и реализованы методы для работы с малыми выборками. Решена задача восстановления функции плотности распределения вероятности и проведено сравнение реализации методов на модельных данных и реальных данных. Помимо этого в работе посчитана

среднеквадратичная ошибка для модельных данных и решена практическая задача о разделении на группы торговых представителей Сибирской кондитерской компании.

Список использованных источников

1. Акимов, С. С. Методы решения задачи восстановления плотности вероятности по выборке из генеральной совокупности / С. С. Акимов // Естетсвенные и математические науки в современном мире. — 2014. — №14. — С. 1–8.
2. Браништи, В. В. Методы и алгоритмы настройки проекционной оценки плотности вероятности случайного вектора в условиях малых выборок: специальность 05.13.17 "Теоретические основы информатики": Диссертация на соискание ученой степени кандидата физико-математических наук / Браништи Владислав Владимирович; Сибирский государственный университет науки и технологий имени академика М.Ф.Решетнёва. — Красноярск, 2018.—125 с.
3. Гаскаров, Д. В. Малая выборка: учебное пособие / Д. В. Гаскаров , В. И. Шаповалов.— М.: Статистика, 1978. — 248 с.
4. Горбунова, Е. Б. Метод статистической обработки малых выборок данных в задачах прогнозирования и контроля состояния сложных систем: специальность 05.13.01 "Системный анализ, управление и обработка информации (техника и технологии)": Диссертация на соискание ученой степени кандидата технических наук / Горбунова Екатерина Борисовна; Южный федеральный университет. — Таганрог, 2018.—178 с.
5. Карандеев, Д. А. Проблема оценивания плотности вероятности по эмпирическим данным / Д. А. Карандеев, И. М. Эйсымонт // Управление большими системами. — 1998.— №1. — С. 48–57.
6. Квишевская, А. Что такое KPI простыми словами / А. Квишевская // Комьюнити: [сайт]. — 2021. — 7 сент. — URL: https://timeweb.com/ru/community/articles/chto-takoe-kpi-prostymi-slovami (дата обращения: 25.06.2023).
7. Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников / А. И. Кобзарь; — М.: ФИЗМАТЛИТ, 2006. — 816 с.

8. Колмогоров, А. Н. Три подхода к определению понятия количество информации / А. Н. Колмогоров // Проблемы передачи информации. — 1965. —Т.1, №1. — С. 3–11.
9. Солонин, С. И. Метод гистограмм: электронное текстовое издание / С. И. Солонин ; Уральский федеральный университет имени первого Президента России Б. Н. Ельцина. — Екатеринбург : УрФУ, — 2014. — 98 с.
10. Усков, В. И. Решение одного интегрального уравнения Фредгольма первого рода / В. И. Усков, В. И. Небольсина // Молодой ученый. —– 2019. —– № 40. —– С. 1-4.
11. Ченцов, Н. Н. Оценка неизвестной плотности распределения по наблюдениям / Н. Н. Ченцов // Доклады Академии наук. — 1962. — Т.147, №1. — С. 45–48.
12. Rosenblatt, M. Remarks on some nonparametric estimates of a density function / M. Rosenblatt // The Annals of Mathematical Statistics.—-1956. – Vol. 27, no. 3. – P. 832–837.

Министерство науки и высшего образования РФ

Федеральное государственное автономное
образовательное учреждение высшего образования

«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики и фундаментальной информатики
Кафедра высшей и прикладной математики

УТВЕРЖДАЮ

Заведующий кафедрой

_____/ С.Г. Мысливец

«22» июня 2023 г.

БАКАЛАВРСКАЯ РАБОТА

Направление 01.03.02 Прикладная математика и информатика

МЕТОДЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ МАЛЫХ ВЫБОРОК ДАННЫХ

Руководитель _____ 22.06.23 доцент, кандидат физико-математических наук Д.В. Семенова

Выпускник _____ 22.06.23 Д.И. Ильина

Нормоконтролер _____ 23.06.23 Т.Н. Шипина

Красноярск 2023