Тема: Машинная обработка статистических данных

Ключевые слова: Статистика, Алгоритмы, Excel, VBA, Машинная обработка, Машинное обучение, Классификация данных.

Аннотация: В данной работе разобран простой алгоритм машинного обучения с учителем для решения задач классификации. Затронуты проблемы, связанные с машинным обучением, реализован и протестирован алгоритм в среде ms excel. В качестве метода классификации был взят алгоритм К ближайших соседей, который отлично подходит для новичков в машинном обучении.

Актуальность: Учитывая доступные вычислительные мощности на данный момент времени машинная обработка данных становится все более востребованной во многих сферах деятельности человека. Использующийся метод k-ближайших соседей является одним из самых простых, актуальность метода обусловлена многочисленностью задач, суть метода заключается в нахождении подходящего решения там, где нет четко сформулированного правила или метода. Область применения данного метода является достаточно широкой, начиная от спам-фильтра почтовых ящиков, заканчивая предсказанием месторождений полезных ископаемых.

Цель работы: Изучить простые алгоритмы классификации данных на прикладном уровне и реализовать алгоритм К ближайших соседей. В процессе проверки алгоритма определить его точность в зависимости от исходных данных и параметра К.

Задачи:

1. Изучить на прикладном уровне простые алгоритмы машинного обучения с учителем и реализовать его для решения задач классификации;
2. Реализовать алгоритм обучения и классификации с различными видами данных с контролем его точности;
3. Внедрить алгоритм в ms excel в виде макросов и/или надстройки;

Материалы и методики: В качестве материалов используется две открытые базы данных. Одна по качеству вина от 1 до 10 [3] и вторая по трем видам ирисов [4]. На основе двух баз данных проводится машинное обучение с учителем и классификация данных по методу К ближайших соседей.

Результаты: Результатом данной работы является алгоритм в среде MS Excel, написанный на языке VBA. Исходный код данного алгоритма представлен в открытом доступе [5], где каждый человек может его модифицировать под себя и предложить эти изменения для общего доступа.

Заключение: На основании реализованного алгоритма были выявлены необходимые требования к базе данных для обучения и тестирования, определена его точность и проблемы, которые могут повлиять на процесс классификации новых данных.

Библиографический список:

1. В. М. Неделько Основы статистических методов машинного обучения [учебное пособие] / В. М. Неделько. – Новосибирск, 2010 – 79 с.
2. А. М. Миронов Машинное обучение / А. М. Миронов. – Москва, 2018 – 90 с.
3. Wine quality data set [Электронный ресурс]: URL: archive.ics.uci.edu/ml/datasets/Wine%20Quality (дата обращения: 03.03.2020)
4. Iris data set [Электронный ресурс]: URL: archive.ics.uci.edu/ml/datasets/Iris (дата обращения: 12.03.2020)
5. Исходный код реализованного алгоритма [Электронный ресурс]: URL: github.com/br0nebr0/VBA\_ML (дата обращения: 21.03.2020)