УДК/ББК ?

**МАШИННАЯ ОБРАБОТКА СТАТИСТИЧЕСКИХ ДАННЫХ**

***Гречихин Андрей Геннадиевич****, студент 3 курса института экономики и управления АПК, РГАУ-МСХА им. К. А. Тимирязева, grechihin.andrey@gmail.com*

***Мелешенков Максим Денисович****, студент 3 курса института экономики и управления АПК, РГАУ-МСХА им. К. А. Тимирязева, rdjdis@yandex.ru*

***Научный руководитель: Романцева Юлия Николаевна****, Доцент РГАУ-МСХА им. К. А. Тимирязева, romantceva@rgau-msha.ru*

***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.*

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

Одна из задач, которая часто встречается в статистике – группировка наблюдений по определенному признаку. Одним из способов решения данной задачи является визуальная группировка, где в качестве осей выступают значения признаков, а сами точки на плоскости наблюдениями, которые необходимо отнести к какой-либо группе.

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.

Алгоритм К ближайших соседей относится к наиболее простым методам машинного обучения с учителем, алгоритм прост для понимания человека и реализации на многих языках программирования. Обучение с учителем подразумевает под собой наличие уже готовой базы знаний, на которую будет опираться алгоритм для принятия новых решений, в данном случае классификации новых наблюдений. [1]

Суть алгоритма сводится к измерению расстояния от нового наблюдения до уже известных, затем выбирают из них К ближайших. На основании результативных признаков К соседей путем голосования определяется результативный признак нового наблюдения. Голосование представляет собой подсчет количества представителей каждой группы среди К ближайших соседей и присоединения наблюдения к большинству [2]. Визуальная схема алгоритма представлена на рисунке 1.

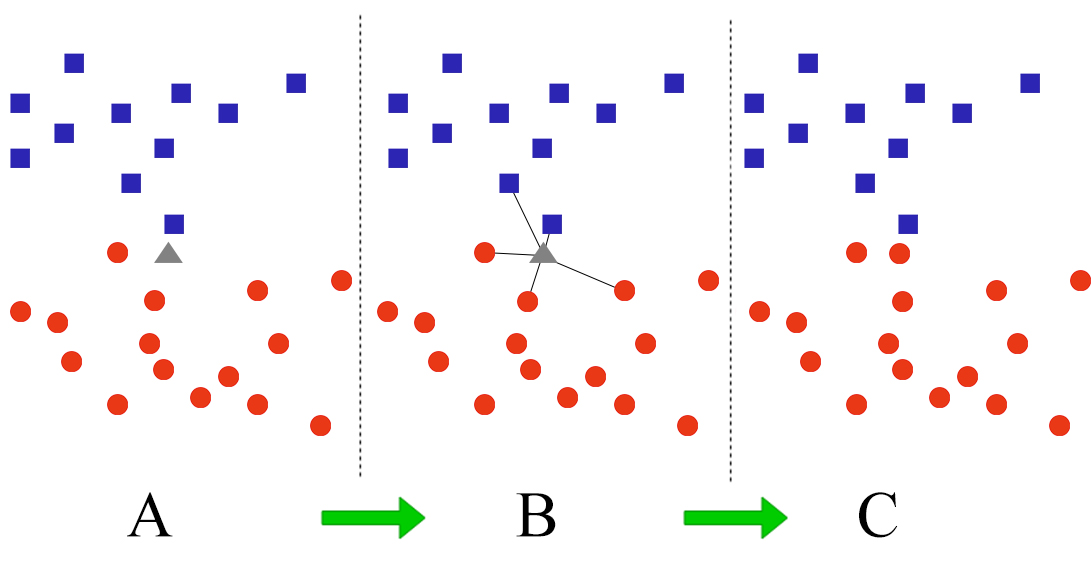


Рисунок 1 - **схема работы алгоритма К ближайшего соседа**

Для проверки работы алгоритма были выбраны две базы данных из открытого банка, которые отличаются по своей сложности классификации.

Первой базой данных являются данные о 3 видах ирисов с 4 признаками (состоит из 150 наблюдений) [3]. Эти данные можно охарактеризовать равной шкалой измерения признаков, что позволяет сразу же позволяет классифицировать эти данные.

Второй базой данных является набор красных вин, которым присваивается оценка от 1 до 10 в зависимости от 11 различных признаков (состоит из 1600 наблюдений). Этот набор данных был использован дважды, во втором случае оценки по десятибалльной шкале были переведены в пятибалльную шкалу оценок.

Основной проблемой классификации вин является неравная шкала измерения признаков, некоторые значения измерялись в сотых, а некоторые в десятках. Для приведения данных значений к единому стандарту была использована Z – оценка, которая благоприятно повлияла на дальнейшую работу алгоритма.

Второй проблемой является наложение групп, встречаются наблюдения, которые располагаются внутри одних групп, но относятся к другим. Для решения этой проблемы в базе данных необходимо отсеять нетипичные наблюдения. Другим способом решения является отбор лишь нескольких наиболее важных признаков. В данном алгоритме эти действия опущены для упрощения понимания его работы.

На основании баз данных были проведены вычисления, зависящие от значения К и применения нормализации при решении задачи классификации (таблица).

Таблица 1

**Результат работы алгоритма**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| RК\* | Точность классификации ирисов, % | | Точность классификации вина, %  (10 баллов) | | Точность классификации вина, %  (5 баллов) | |
| с нормализацией | без нормализации | с нормализацией | без нормализации | с нормализацией | без нормализации |
| 1 | 86,3 | 96 | 43,3 | 42 | 65,3 | 74,5 |
| 2 | 92,2 | 94 | 46,5 | 41 | 73,3 | 75,7 |
| 3 | 92,2 | 96 | 51 | 40,2 | 77,8 | 79 |
| 4 | 88,2 | 94 | 47,8 | 37,8 | 78,3 | 75,8 |
| 5 | 90,2 | 94 | 49,7 | 39,7 | 78,7 | 79,2 |
| 6 | 84,3 | 90 | 53,3 | 43,5 | 79 | 77,7 |
| 7 | 90,2 | 90 | 50,3 | 44 | 80,8 | 79,3 |
| 8 | 78,4 | 90 | 53,2 | 42,7 | 81 | 79 |
| 9 | 84,3 | 90 | 52,3 | 42,5 | 82 | 79,3 |
| 10 | 82,4 | 88 | 51,3 | 43,8 | 81,8 | 79,8 |

Исходя из данных таблицы можно сказать, что нормализация улучшает результат только при неоднородных данных параметров, к примеру точность с нормализацией у вин в целом выше, чем без нее.

Также можно заметить, что точность зависит от К не напрямую, а это означает, что оптимальное значение К необходимо подбирать для каждого набора данных индивидуально методом перебора.

По сути перебором занимается реализованный алгоритм [5] при обучении, после чего с определенной точностью он может классифицировать новые наблюдения.

**Библиографический список**

1. А. М. Миронов Машинное обучение / А. М. Миронов. – Москва, 2018 – 90 с.
2. В. М. Неделько Основы статистических методов машинного обучения [учебное пособие] / В. М. Неделько. – Новосибирск, 2010 – 79 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)