

ISSN 1815-4840



КИМЁВИЙ ТЕХНОЛОГИЯ НАЗОРАТ ВА БОШҚАРУВ

Халқаро илмий-техникавий журнал 2/2014

ISSN 1815-4840

ТОШКЕНТ ДАВЛАТ
ТЕХНИКА УНИВЕРСИТЕТИ,
«ЎЗҚУРИЛИШМАТЕРИАЛЛАР» ДАҚ,
«ЎЗКИМЁСАНОАТ» ДАҚ,
«СОВПЛАСТИТАЛ» ҚҚ,
ЎЗБЕКИСТОН РЕСПУБЛИКАСИ
ИНТЕЛЛЕКТУАЛ МУЛК АГЕНТЛИГИ

**КИМЁВИЙ ТЕХНОЛОГИЯ.
НАЗОРАТ ВА БОШҚАРУВ**

2014, №2(56)

Халқаро
илмий - техникавий журнал

2005 йилдан нашр этилади



ХИМИЧЕСКАЯ ТЕХНОЛОГИЯ. КОНТРОЛЬ И УПРАВЛЕНИЕ

2014, №2(56)
Международный
научно-технический журнал

Издается с 2005 года

Учредители:

ТАШКЕНТСКИЙ ГОСУДАРСТВЕННЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ,
ГАО «УЗСТРОЙМАТЕРИАЛЫ»,
ГАО «УЗКИМЕСАНОАТ»,
СП «СОВПЛАСТИТАЛ»,
АГЕНТСТВО ПО ИНТЕЛЛЕКТУАЛЬНОЙ
СОБСТВЕННОСТИ РЕСПУБЛИКИ
УЗБЕКИСТАН

Главный редактор и председатель
редакционной коллегии:
ЮСУПБЕКОВ Надырбек Рустамбекович

Заместители:

АЗИМОВ Ахил Адылович,
ГУЛЯМОВ Шухрат Маннапович,
ИГАМБЕРДИЕВ Хусан Закирович

Ответственный секретарь:
МАННАНОВ Уллубек Васикович

Редакционная коллегия:

Азимов Р.К., Акрамов Э.М., Алиев Р.А.
(Азербайджан), Аллаев К.Р., Балакирев В.С.
(Россия), Бекмуратов Т.Ф., Беглов Б.М.,
Бишимбаев В.К. (Казахстан), Верлань А.Ф.
(Украина), Гордеев Л.С. (Россия), Глушенкова
А.И., Закиров С.Г., Ибрагимов Г.И., Исмаилов
М.А., Исмаилов П.Р., Каландаров П.И.,
Марохинов А.Р., Мухамедханов У.Т., Камбаров
М.М., Касымов С.С., Кузнецова Н.Н. (Россия),
Ладанюк А.П. (Украина), Мамаджанов Х.А.
(Россия), Мещалкин В.П. (Россия), Мелкумов
А.Н., Мирзаалимов М.С., Мухитдинов М.М.,
Мухитдинов Д.Н., Нурмухамедов Х.С.,
Набиев О.М., Назаров У.С., Рашидова С.Ш.,
Сайдахмедов Р.Х., Салимов З.С., Ульянов С.В.
(Россия), Усманов Р.Н., Хакимов О.Ш.,
Чистякова Т.Б. (Россия),
Юсупбеков А.Н., Якубов Р.Я.

Адрес редакции:

100095, г. Ташкент
ул. Университетская, 2

Телефон:

227-17-16

e-mail:

app-igti@mail.ru



Материалы, публикуемые в настоящем журнале, не
могут быть полностью или частично воспроизведены без
письменного разрешения редакции. Мнение редакции не
всегда совпадает с мнением авторов материалов. За
достоверность сведений, представленных в журнале,
ответственность несут авторы статей и редакторы.

СОДЕРЖАНИЕ

ХИМИЧЕСКАЯ ТЕХНОЛОГИЯ

- К.А.Режаббаев, У.К.Алимов, Ш.С.Наматов, Б.М.Беглов.
Моноаммонийфосфат и азотнофосфорнокислый серосодержащие
удобрения на основе фосфорнокислотного разложения небогатых
фосфоритов Центральных Кызылкумов 5
- Т.А.Атакузиев, Г.А.Тохтагулова, Г.Ж.Оразимбетов, Влияние
добавок NaCl , CaCl_2 на процесс минералообразования при обжиге
сульфатсодержащих шихт 12
- О.Рахматов, Г.Р.Иргашева, Энергосберегающая технология
бланширования винограда 17
- И.Н.Турсунова, У.М.Марлонов, М.Н.Муратова, У.А.Эркаев.
Использование продуктов переработки низкосортных фосфоритов в
производстве азотно-фосфорных удобрений 22

КОНТРОЛЬ ТЕХНОЛОГИЧЕСКИХ ПАРАМЕТРОВ

- С.Ф.Амиров, А.М.Сафаров, Д.Ш.Рустамов, Электромагнитный
датчик тока для систем управления устройствами тягового
электрооборудования 26
- Р.К.Алимов, Ю.Г.Шипулин, О.С.Райимжанова, С.А.Максудов.
Датчики на основе сегментных фоторезисторов для контроля угловых
перемещений 31
- Г.И.Зарипова, Методы адаптивного контроля достоверности
обработки данных непрерывных объектов на основе нейро-нечетких
моделей 36
- Илхомжон Х.Сиддинов, О.Г.Салиев, Ж.И.Сиддинов,
З.Ж.Самиев, Алгоритмы структурного и параметрического
проектирования электромагнитных преобразователей 42

УПРАВЛЕНИЕ ТЕХНОЛОГИЧЕСКИМИ ПРОЦЕССАМИ

- М.М.Арипов, Д.К.Мухамедиева, Популяционные модели типа
Колмогорова-Физера с нелинейной кросс-диффузией 46
- И.И.Жуманов, Н.А.Шаранова, Системы анализа и обработки данных
о кадрах вузов на основе нечеткого регулятора с фиксированной базой
знаний 53
- Ф.Т.Ахмедов, С.В.Жоров, Н.И.Турапина, Advanced process control –
усовершенствованное управление технологическими процессами и
производствами 61
- С.К.Ганиев, О.У.Халмуратов, З.Т.Худойкулов, Определение
весовых коэффициентов классов функциональных требований
стандарта «Методы и средства обеспечения безопасности. Критерии
оценки безопасности. Информационные технологии» 68
- Д.Т.Мухамедиева, Применение мягких вычислений при построении
моделей задач принятия решений по оценке состояния
слабоформализуемого процесса 73
- О.Б.Рузибаев, О.К.Хужаев, Исследование и программная реализация
метода ближайших соседей 84
- С.Т.Юнусова, Формализация информационной модели процесса
переработки хлопчатника 90

УДК 004.4

О.Б.РУЗИБАЕВ, О.К.ХУЖАЕВ (ГУИТ)

ИССЛЕДОВАНИЕ И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ МЕТОДА БЛИЖАЙШИХ СОСЕДЕЙ

Хозиреи вақтда тимсолларни аниқлаш масаласи кўп амалий соҳалардаги долзарб масалалардан бири ҳисобланади. Тимсолларни аниқлашда синфлаш усулларида кенг фойдаланилади, уларнинг ичида энг кенг тарқалгани К-яқин қўнилар (KNN) усулидир. Бу усул башоратлаш масаласини ҳам ечишга имкон беради. Шунинг учун усул асосида амалий масалаларни ечишни автоматлаштирувчи дастурий таъминот яратилиши муҳим аҳамиятга эга. Шунингдек назарий жиҳатдан хатоликларни таҳлил қилиш ва оптимал параметрларни танлашни автоматлаштириш ҳам муҳимдир. Ниҳода баъзи амалий масалалар учун усул хатолиги келтирилган ва усулни оптимал қўллашга тавсиялар келтирилган. Объектга йўналтирилган дастурлаш технологияси асосида яратилган дастурий таъминот тавсифи келтирилган.

Таянч сўзлар: тимсолларни аниқлаш, таснифлаш, регрессия, К-яқин қўнилар усули, Канберра, Соренсен, ҳамма вариантларни кўриб чиқиш, Байеснинг таянч векторлар усули.

Распознавание образов в настоящее время является одной из актуальных задач в различных практических областях. При распознавании образов широко применяются методы классификации, среди которых наиболее распространенным является метод К-ближайших соседей (KNN). Данный метод также позволяет решать задачи прогнозирования. Поэтому разработка программного обеспечения для автоматизации решения практических задач на основе данного метода является злободневной. Актуальной с теоретической точки зрения является и автоматизация исследования погрешностей и выбор оптимальных параметров исследуемых объектов. В работе для практических задач проведен анализ погрешностей метода, а также даны практические рекомендации по оптимальному их применению. Дано описание программного обеспечения, созданного на основе объектно-ориентированной технологии программирования.

Ключевые слова: распознавание образов, классификация, регрессия, метод К-ближайших соседей, Канберра, Соренсена, перебор, метод опорных векторов, Байеса.

Pattern recognition is currently one of the important tasks in various practice area. Classification methods are widely used in pattern recognition. K nearest neighbors (KNN) is the most common method among these methods. This method also allows you to decide the problem of forecasting. Therefore, the development of software to automate the solution of practical problems on the basis of this method is relevant. Also relevant from a theoretical point of view is the automation trials errors and choice of optimal parameters. In this paper, we analyzed errors of the methods, as well as practical recommendations for its optimal application. Also describes the software created on the basis of object-oriented programming techniques.

Key words: Pattern recognition, classification, regression, method K-nearest neighbors, Canberra, Sorenson, brute force, support vector machine, Bayes.

1. Введение. Существует значительное число методов классификации – таких, как К-ближайших соседей, нейронной сети, деревья решений, Байеса, метод опорных векторов (SVM). Среди них метод К-ближайших соседей KNN (k-nearest neighbor) является наиболее популярным методом классификации, демонстрируя хорошие характеристики и короткий период времени обучения. Метод К-ближайших соседей является простым, но очень эффективным и действенным в области распознавания образов, категоризации текста, медицине и т.д.

Идея данного метода состоит в следующем: зная все объекты в обучающей выборке, алгоритм находит К-ближайших объектов к неизвестному объекту в обучающей выборке. Неизвестный объект считается принадлежащим классу, к которому принадлежит большинство найденных объектов. Данный метод может применяться для регрессии, только при этом в качестве результата определяется среднее К-ближайших точек [1].

Одна из проблем метода заключается в том, что при классификации решение может быть неоднозначным. Поэтому при разделении на два класса К выбирается нечетным. Но если число классов больше даже при нечетном К, результат может быть неоднозначным. При решении задачи регрессии данная проблема не возникает, так как вычисляются средние значения. Определение оптимального К является практически важной задачей. Обычно

алгоритм KNN работает точно, но все алгоритмы классификации и прогнозирования имеют погрешности, и поэтому в статье проведен анализ погрешностей метода для различных задач. Для алгоритма KNN важно определение оптимального значения параметра K. Если K слишком мало, алгоритм может не охватить нужную информацию; если слишком большое, может быть получено слишком много ненужной информации и в результате получить неправильный прогноз.

До настоящего времени анализ погрешностей метода KNN достаточно не исследован. Большинство работ посвящено сравнению метода с альтернативными алгоритмами [2].

2. Этапы и оптимизация метода.

Стандартизация. Один из наиболее распространенных методов при обработке данных – приведение к стандартному виду. В статистике стандартизация может быть осуществлена на основе следующей формулы:

$$z = \frac{x - \mu}{\sigma}.$$

Здесь: μ – среднее значение σ , $(\sigma - x)$ – отклонение от среднего.

$$\mu = \frac{\sum_{i=1}^n x_i}{n}, \quad \sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}}, \quad n - \text{объем выборки}.$$

После стандартизации x приводится к распределению со средним значением 0 и с дисперсией 1.

Нормализация признаков. Один из распространенных методов приведения к стандартному виду – нормализация:

$$z = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

Результат после нормализации лежит в промежутке $0 \leq z \leq 1$.

Выбор метрики. При решении задачи классификации и регрессии методом K – ближайших соседей наиболее часто используются метрики Евклида, Манхеттена, Чебышева, Минковского, Канберра, Соренсена, а также метод углов.

Метрика Евклида. Метрика Евклида основана на формуле Пифагора:

$$d_{ij} = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}. \quad (1)$$

Здесь, n – число параметров.

Метрика Манхеттен. В этой метрике вычисляется сумма абсолютных значений разностей признаков:

$$d_{ij} = \sum_{k=1}^n |x_{ik} - x_{jk}|. \quad (2)$$

Метрика Чебышева. В метрике определяется максимум абсолютных значений разностей признаков, поэтому данный метод называется методом максимального значения:

$$d_{ij} = \max_k |x_{ik} - x_{jk}|. \quad (3)$$

Метрика Минковского. Метрика Минковского считается обобщенной, поскольку при $\lambda=1$ превращается в метрику Манхэттена, а при $\lambda=2$ – в метрику Евклида; при $\lambda=\infty$ она превращается в метрику Чебышева:

$$d_{ij} = \sum_{k=1}^n |x_{ik} - x_{jk}|^{\lambda}. \quad (4)$$

Метрика Канберра. Для этой метрики нормализация необязательна, так как для каждого параметра значение лежит в промежутке $[0,1]$, но сумма в этом промежутке может не лежать. Недостаток данного метода в том, что оба параметра не должны быть равны нулю:

$$d_{ij} = \sum_{k=1}^n \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}. \quad (5)$$

Метрика Соренсена. Метрика Соренсена также используется при нормализации данных. Если значения параметров положительны, то результат лежит в промежутке $[0,1]$. Недостаток данного метода состоит в том, что оба параметра не должны быть равны нулю:

$$d_{ij} = \frac{\sum_{k=1}^n |x_{ik} - x_{jk}|}{\sum_{k=1}^n (x_{ik} + x_{jk})}. \quad (6)$$

Метод углов. Вычисляется косинус угла между двумя векторами. Метод направлен на определение степени близости векторов. Здесь не требуется нормализация данных, так как косинус угла всегда лежит в промежутке $[-1, +1]$. В программе можно вычислить К минимальных расстояния. Поэтому в программе используется $\frac{1}{\cos \alpha}$.

$$d_{ij} = \frac{\sum_{k=1}^n x_{ik} * x_{jk}}{\sqrt{\sum_{k=1}^n x_{ik}^2 * \sum_{k=1}^n x_{jk}^2}}. \quad (7)$$

3. Оптимизация метода KNN. Практическое применение принципа KNN требует сложных алгоритмов.

Для оптимизации метода KNN в программе использован метод "brute force" для определения ближайших объектов. Недостаток метода состоит в том, что требуется много времени и большой объем памяти. Поэтому для оптимизации была использована структура данных **Priority Queue** (приоритетная очередь). Особенность данной структуры в том, что максимальный элемент всегда передвигается в начало. Вставка и чтение элемента производятся с логарифмической сложностью, а не квадратичной, как обычно.

Обычно для определения уровня погрешности проводится несколько тестов и средняя погрешность принимается как истинная. Обычно для выбора оптимального K используется **K-fold-cross-validation** метод. Проводится 100 тестов и средняя погрешность принимается как истинная погрешность. Каждый раз при проведении теста 90% используется для изучения точки и 10% используется для определения погрешности. При каждом тесте места точек меняются. При перестановке для генерации случайных индексов используется алгоритм **Кнута**.

4. Анализ полученных результатов. Для обучения и классификации, а также для сравнения погрешности прогнозирования были использованы данные из сайта Калифорнийского Университета archive.ics.uci.edu, обработанные с помощью разработанного программного обеспечения. В таблице приведен список использованных материалов [3]:

Набор данных	Число объектов	Количество признаков
Автомобиль	398	8
Рак груди	699	10

В этой задаче на основе множества данных «автомобиль» была решена задача прогноза мощности в лошадиных силах на основе других параметров. Результаты решения данной задачи для различных метрик и определения оптимального K приведены в табл. 1.

Из таблицы видно что для множества данных «автомобиль» метрика Манхеттена при $K=8$ и метрика Канберра при $K=4$ даёт наименьшую погрешность. При использовании метрики Манхеттена параметр стандартизируется, а для метрики Канберра эту задачу выполняет знаменатель, числители одинаковы. Необходимо отметить что при использовании метрики Минковского с увеличением λ погрешность увеличивается. Погрешность метрики Чебышева больше метрики Минковского всех степеней.

Таблица 1

Набор данных «автомобиль»

K	1	2	3	4	5	6	7	8	9	10	Оптимальный K	Самые меньшие ошибки
Евклид	9.17	8.49	8.09	8.01	8.1	8.1	8.03	7.99	8.11	8.15	8	7.99
Манхеттон	9.42	8.23	7.97	8.1	8.07	7.86	8.03	7.85	7.86	7.9	8	7.85
Чебышев	9.31	8.69	8.46	8.22	8.15	8.45	8.26	8.26	8.27	8.51	5	8.15
Минковский $\lambda=3$	9.2	8.5	8.07	8.02	8.26	7.97	8.11	7.92	8.03	8.02	8	7.92
Минковский $\lambda=4$	9.45	8.4	8.04	8.26	8.32	8.25	8.23	8.12	7.94	8.2	9	7.94
Минковский $\lambda=5$	9.46	8.29	8.27	8.25	8.34	8.13	8.11	8.05	8.17	8.33	8	8.05
Канберра	8.23	8.17	7.99	7.85	7.95	8.05	8.01	7.9	8.01	8.04	4	7.85
Соренсон	12.4	11.8	11.1	11.1	11.1	11.1	10.9	11	11	11.3	7	10.94
Бурчак	10.8	10.6	10	10	10	9.79	9.79	9.79	9.94	10.3	8	9.79

Если для множества данных «рак груди» рассмотрена задача регрессии, то для задачи базы данных «грудиной рак» решена задача классификации. Результаты решения приведены в табл. 2.

Таблица 2

Набор данных «Рак груди»

K=	1	2	3	4	5	6	7	8	9	10	Оптимальный K	Самые меньшие ошибки
Евклид	4.36	4.06	3.91	3.38	3.54	3.03	2.77	3.39	3.46	2.96	7	2.77
Манхеттон	3.16	3.72	3.3	3.61	3.36	3.59	3.8	3.58	3.43	3.16	10	3.16
Чебышев	4.87	5.52	4.29	4.1	3.78	3.59	3.77	3.49	3.58	3.25	10	3.25
Минковский $\lambda=3$	5.06	4.74	3.62	3.3	3.36	3.51	3.41	3.58	3.22	3.33	9	3.22
Минковский $\lambda=4$	5.26	4.91	3.68	3.83	2.83	3.75	3.36	3.49	3.3	3.81	5	2.83
Минковский $\lambda=5$	5.38	4.52	3.51	3.7	3.26	3.43	3.7	3.81	3.57	3.54	5	3.26
Канберра	3.87	3.91	2.7	3.03	2.62	3.33	2.86	3.64	2.86	3.23	5	2.62
Соренсон	2.84	3.7	2.38	2.67	2.48	2.62	2.8	2.65	2.97	3.06	3	2.38
Бурчак	10.5	11.3	10.3	11.1	11	10.9	9.94	9.1	9.17	9.51	8	9.1

Из таблицы видно, что метрика Соренсона и использование трех наиболее близких точек дает наименьшую погрешность при разделении на два класса.

5. Структура программного обеспечения. Программное обеспечение создано на языке C# и имеет удобный для пользователя интерфейс. Оно состоит из классов для реализации алгоритмов классификации и регрессии, а также online форм для отражения результатов.

Интерфейс пользователя создан в среде Microsoft Visual Studio 2010 и состоит из нескольких окон. В верхней части основного окна расположено меню, которое состоит из разделов: выбор файла, результат, гистограмма, уровень погрешности, график, оптимальное K и сведения о программе. В нижней части отражается информация о готовности решения [4].

В правой части основного окна можно выбрать метрику и ввести значение K, количество тестов, точность и процент обучения. В левой части основного окна выбираются атрибуты и тип задачи. Центральная часть основного окна предназначена для отражения полученных результатов. На рис.1 приведено решение задачи регрессии для автомобиля и уровень погрешности методом KNN. На рис.2 приведено решение методом KNN задачи выбора оптимального K для автомобиля.

На рис.3 отражены результаты решения задачи классификации для базы данных «рак груди».

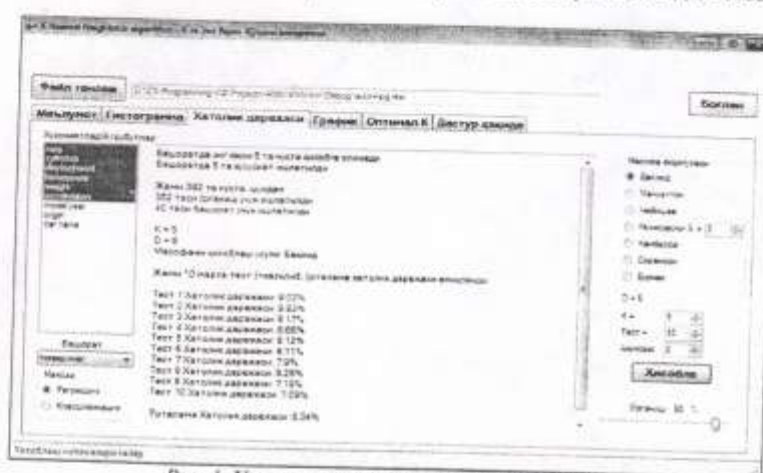


Рис. 1. Уровень погрешности для автомобиля.

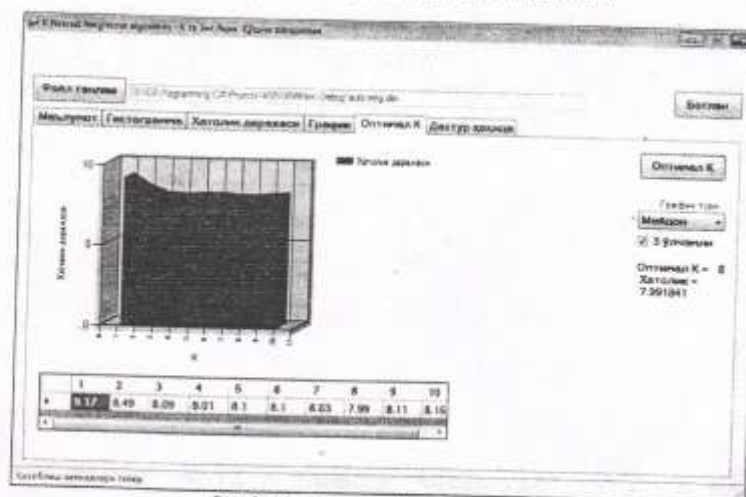


Рис. 2. Множество данных «автомобиль».

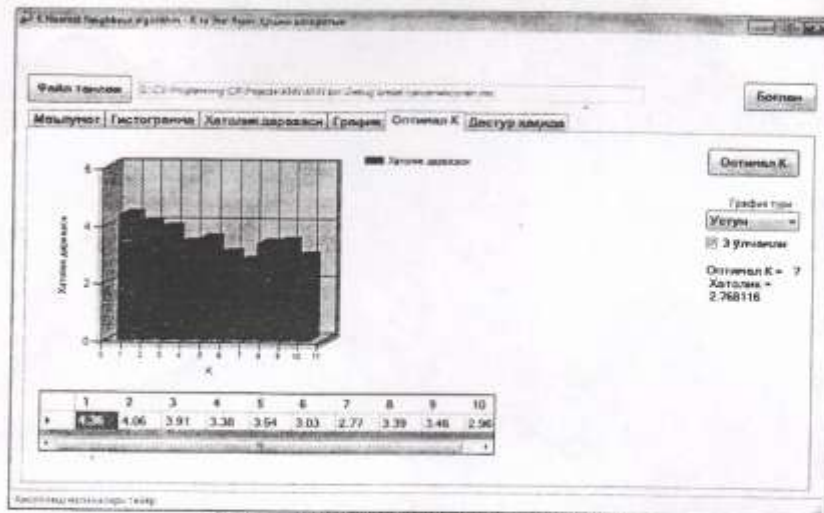


Рис. 3. Множество данных задачи «рак груди».

Выводы

Полученные результаты показывают, что при решении задачи регрессии наименьшая погрешность (7.85%) получена, когда использована метрика Манхеттена при $K=4$. При решении задачи классификации грудного рака наименьшая погрешность (2.38%) получена, когда использована метрика Соренсена при $K=3$. Необходимо отметить, что в статистике модели, погрешности которых меньше 5%, считаются правильными. А это значит, что в дальнейшем при решении задачи классификации рака груди можно использовать $K=3$ и метрику Соренсена.

Можно отметить, что алгоритм KNN даёт хорошие результаты при решении практических задач регрессии или классификации.

Список литературы:

1. Shamsul Huda Md., K. Md. R. Alam, K. Mutsuddi, Md. K. S. Rahman and C. M. Rahman, "A Dynamic K-Nearest Neighbor Algorithm For Pattern Analysis Problem", 3rd International Conference on Electrical & Computer Engineering, ICECE 2004, Dhaka, Bangladesh, 28-30 December 2004.
2. James T. O. "Comparison Classifier of Condensed KNN and K-Nearest Neighborhood Error Rate Method", The computing science and technology international journal, vol. 2, no. 3, september 2012.
3. Asuncion A. and D.J. Newman: "UCI Machine Learning Repository", Irvine, CA: University of California, School of Information and Computer Science, 2007. [Online]. Available: <http://archive.ics.uci.edu/ml/datasets.html>.
4. Троелсен, Э., C# и платформа .NET. Библиотека программиста. — СПб.: Питер, 2004. — 796 с.: ил., ISBN 5-318-00750-3.
5. Zhan Y., H. Chen and G. C. Zhang, "An Optimization Algorithm Of K-NN Classification", Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006.

Рүзиев Ортик Бахтиёрович – старший научный сотрудник-соискатель кафедры «Программное обеспечение информационных технологий» ТУИТ, Тел.: (+99891)166-17-47; E-mail: ortiq231986@mail.ru.

Хужаев Отабек Қадамбоевич – старший научный сотрудник-соискатель кафедры «Программное обеспечение информационных технологий» ТУИТ, Тел.: (+99891)-913-30-50; E-mail: otabek.hujayev@gmail.com.