ISSN 2181-4732
ISSN 2181-4732

УДК: 62-051.31:62-051.15

КЛАССИФИКАЦИЯ АГРАРНОЙ ЗЕРНОВОЙ ПРОДУКЦИИ НА ЗДОРОВЫЕ И ЗАРАЖЕННЫЕ ПЛЕСЕНЬЮ ПО ИХ ЦВЕТОВЫМ ХАРАКТЕРИСТИКАМ

Мухамедханов Улугбек Тургудович[1]- доктор технических наук, профессор,
ORCID: 0009-0001-7579-1885, E-mail: m-ulugbek@mail.ru
Сувонов Бехруз Искандар угли[2]- старший преподаватель
ORCID: 0009-0002-4111-6889, E-mail: suvonovbekhruz@gmail.com

[1]Ташкентский государственный технический университет
имени Ислама Каримова, г. Ташкент, Узбекистан
[2]Университет экономики и педагогики, г. Карши, Узбекистан

Аннотация. *В данной статье представлены различные этапы построения системы классификации аграрной зерновой продукции на здоровые и заражённые плесенью. На первом этапе рассматриваются и тестируются методы выбора подмножеств цветовых признаков зерновой продукции. Селектируются те из них, которые обладают высокой дискриминационной информативностью относительно задачи классификации. Затем к отобранной продукции применяются различные классификаторы, производительность и точность которых оценивается и выбирается.*

Ключевые слова: *распознавание образов, системы классификации и идентификации, цветовые признаки зерновой продукции.*

UDC: 62-051.31:62-051.15

CLASSIFICATION OF AGRICULTURAL GRAIN PRODUCTS INTO HEALTHY AND MOLD-CONTAMINATED CATEGORIES BASED ON THEIR COLOR CHARACTERISTICS

Mukhamedkhanov, Ulugbek Turgudovich[1] – Doctor of Technical Sciences, Professor
Suvonov, Behruz Iskandarov ugli[2] – Senior Lecturer

[1]Tashkent State Technical University named after Islam Karimov, Tashkent city, Uzbekistan
[2]University of Economics and Pedagogy, Karshi city, Uzbekistan

Abstract. *This article presents various stages of developing a classification system for agricultural grain products into healthy and mold-contaminated categories. In the first stage, methods for selecting subsets of color features of grain products are examined and tested. Those features that possess high discriminative informativeness with respect to the classification task are selected. Then, various classifiers are applied to the selected products, and their performance and accuracy are evaluated and compared to determine the most effective one.*

Keywords: *pattern recognition, classification and identification systems, color features of grain products.*

UO‘K: 62-051.31:62-051.15

AGRAR DON MAHSULOTLARINI ULARNING RANG XUSUSIYATLARI ASOSIDA SOG‘LOM VA MOG‘OR BILAN ZARARLANGAN TURLARGA AJRATISH

Muxamedxanov Ulug‘bek Turg‘unovich[1] – texnika fanlari doktori, professor
Suvonov Behruz Iskandar o‘g‘li[2] – katta o‘qituvchi

[1]Islom Karimov nomidagi Toshkent davlat texnika universiteti, Toshkent sh., O‘zbekiston
[2]Iqtisodiyot va pedagogika universiteti, Qarshi sh., O‘zbekiston

Annotatsiya. *Ushbu maqolada agrar don mahsulotlarini sog‘lom va mog‘or bilan zararlangan toifalarga ajratish uchun klassifikatsiya tizimini qurishning turli bosqichlari bayon ətilgan. Dastlabki bosqichda don mahsulotining rang belgilari (koloristik xususiyatlari) bo‘yicha belgilar*

ISSN 2181-4732
ISSN 2181-4732

to‘plamlaridan kichikroq past toifalarni (submno’jinalarni) tanlash va ularni testdan o‘tkazish usullari ko‘rib chiqiladi. Klassifikatsiya vazifasida yuqori diskriminatsion informativlikka ega bo‘lgan rang belgilarining aniq kombinatsiyalari seleksiya qilinadi. So‘ngra saralab olingan belgilar majmuasiga turli klassifikatorlar qo‘llanilib, ularning ishlash samaradorligi va aniqlik ko‘rsatkichlari baholanadi hamda optimal klassifikator tanlab olinadi.

Kalit so‘zlar: *obrazlarni tanish, klassifikatsiya va identifikatsiya tizimlari, don mahsulotlarining rang belgilari.*

Введение

Частным случаем задачи распознавания образов является классификация объектов по их принадлежности к определённым группам (классам). Классификация - это информационный процесс, заключающийся в преобразовании информации о значениях признаков, описывающих объекты классификации, в информацию об их принадлежности к заранее определённому классу [1].

При выборе адекватного классификатора основными критериями являются применимость, эффективность и время классификации при достаточно высокой точности. Эффективность различных классификаторов существенно зависит от статистических характеристик входных данных, используемых при построении классификатора, то есть от обучающей выборки и априорной информации [2, 3].

Синтез признаков для распознавания и выбор комплекса наиболее информативных признаков являются одними из важнейших этапов при разработке алгоритма распознавания, для которого в большинстве практических задач не существует универсального подхода. Посредством отбора наиболее значимых/информативных признаков ставится цель - визуализировать классы как чётко разделённые области в признаковом пространстве. Другим аспектом, связанным с определением значимости набора признаков, является нахождение наилучшей комбинации признаков, состоящей из минимального количества параметров.

Тематика создания классификационных моделей является объектом исследования ряда авторов [4, 5, 6]. В [7] представлен сравнительный анализ нескольких техник отбора признаков, использующих различные критерии — Information Gain (IG), Gain Ratio (GR), Chi-square (CS), Relief-F. Эффективность этих методов была оценена по точности классификации с использованием классификаторов: k-ближайших соседей (k-NN).

Учитывая изложенное, основной целью настоящей работы является выбор оптимальной классификационной стратегии для определения качества сельскохозяйственной продукции с целью её разделения на два класса: здоровые кукурузные зёрна и заражённые кукурузные зёрна (поражённые плесенью рода *Fusarium Moniliforme* - розовая фузариозная гниль).

Основная часть

Основные направления, которые необходимо учитывать при синтезе классификаторов аграрной продукции по качественным признакам, согласно литературному обзору по данной теме, включают: формализацию описания объекта; формирование обучающей выборки; обучение классификатора; снижение размерности признакового пространства; классификация; определение точности классификации.

При формировании исходного признакового пространства из каждого пикселя изображения зёрен извлекаются цветовые признаки соответствующих компонентов.

База данных содержит значения цветовых компонент цветовой модели.

1. Критерии и методы, используемые для определения диагностической "ценности" признаков для классификации здоровых и заражённых зёрен.

При решении задачи по нахождению оптимального минимального подмножества признаков в настоящем исследовании использованы несколько критериев для определения диагностической значимости, как по отдельности для каждого признака, так и в комбинации.

1.1. Критерии для индивидуальной оценки значимости признаков.

ISSN 2181-4732
ISSN 2181-4732

➢ **FDR (Fisher's Discriminant Ratio)** — критерий будет высоким, и признак будет считаться информативным.

$$\text{FDR} = \frac{(m_1 - m_2)^2}{\sigma_1^2 + \sigma_2^2}, \quad (1)$$

где m_1 - среднее значение данного признака в первом классе; m_2 - среднее значение признака во втором классе; σ_1^2 - дисперсия признака в первом классе; σ_2^2 - дисперсия признака во втором классе.

➢ **Критерий χ^2 (хи-квадрат, Chi-square/CS)** - CS измеряет известную χ^2-статистику для всех значений подмножества входных признаков. Полученная величина χ^2 соответствует степени зависимости между признаком и соответствующим классом. Значение критерия χ^2 для признака F вычисляется по следующей зависимости:

$$\chi^2(F) = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}, \quad E_{ij} = \frac{R_i \cdot C_j}{|S|}, \quad (2)$$

где: m - количество подгрупп признакам F; k - количество классов; A_{ij} - значение признака в i-й подгруппе для j-го класса; E_{ij} - ожидаемое значение признака A_{ij} в i-й подгруппе; C_j - количество признаков в j-м классе; R_i - количество наблюдений в i-й подгруппе; $|S|$ - общее число наблюдений.

1.2. Критерии оценки значимости комбинаций признаков (т.н. векторов признаков – *feature vectors*) и оценки разделимости классов в соответствующем признаковом пространстве

➢ **Критерий J_3** основан на диаграммах рассеивания (*Scatter matrices*):

$$J_3 = \text{trace}\{S_w^{-1} S_b\},$$

где: S_w - внутриклассовая ковариационная матрица; S_b - межклассовая ковариационная матрица.

1.3. Feature Subset Selection - Для достижения рациональной комбинации информативных признаков, необходимо применить оценку всех возможных комбинаций признаков с учётом коэффициента взаимной корреляции между ними (*cross-correlation coefficient*).

➢ **Scalar Feature Selection** — Метод, при котором рассчитывается коэффициент взаимной корреляции между первым (наиболее сильным) признаком и каждым последующим.

➢ **Дискриминантный анализ** (*General Discriminant Analysis*) — Представляет собой концепцию выбора наилучшего подпространства из всех возможных, в котором будет построена классификационная модель. Специальная количественная реализация метода дискриминантного анализа базируется на критерии Уилкса (по Фишеру), один из двух детерминантных критериев — чем он меньше, тем лучше. Он вычисляется как:

$$\Lambda = \frac{\det W}{\det T},$$

где W - внутриклассовая ковариационная матрица; T - общая ковариационная матрица. Критерий также известен как критерий Уилкса или λ-критерий.

Программные инструменты, использованные для проведения анализов, включают интерактивную программную среду MATLAB 7.1 и STATISTICA 8.0 (StatSoft, Inc.).

Получены комбинации информативных признаков для 8 сортов кукурузных зёрен и трёх групп данных: изображения зёрен, снятых только с «гладкой» стороны; изображения, снятые со стороны зародыша; и смешанные базы данных. Для анализа применены три критерия. Критерии Scalar Feature Ranking и Best Feature Combination используют базы данных с нормализованными значениями признаков, тогда как при методе GDA (General Discriminant Analysis) используются необработанные значения признаков, соответствующие компонентам различных цветовых моделей.

В методе GDA общее количество протестированных комбинаций признаков для каждой подвыборки каждого сорта составляет 131 069. Это число полностью определяется количеством признаков в исходном подмножестве, в данном случае - 17 признаков.

2. Сравнительный анализ стандартных методов классификации. Результаты классификации с использованием отобранных комбинаций информативных признаков

ISSN 2181-4732
ISSN 2181-4732

Рассмотрены особенности и характеристики трёх типов классификаторов и проведено сравнительное исследование в задаче распознавания изображений здоровых и заражённых фузариозом кукурузных зёрен.

➢ Классификатор k-NN (k-ближайших соседей) - непараметрический классификатор, основанный на принципе классификации через ассоциацию с группой образов одного класса, называемых ближайшими соседями. Каждый класс представлен совокупностью эталонных образов. Неизвестный объект относится к тому классу, к которому принадлежит по меньшей мере **S** из **k** ближайших соседей из обучающей выборки. Порог достоверности **S** обычно выбирается в пределах: $k/2 < S \leq k$. Мера сходства/близости выбирается в зависимости от задачи; наиболее часто используется евклидово расстояние.

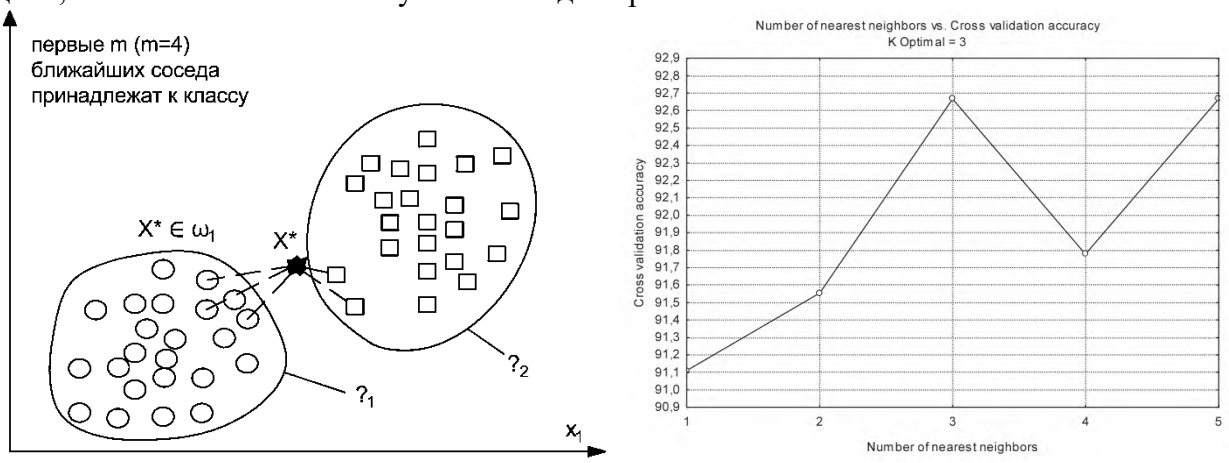
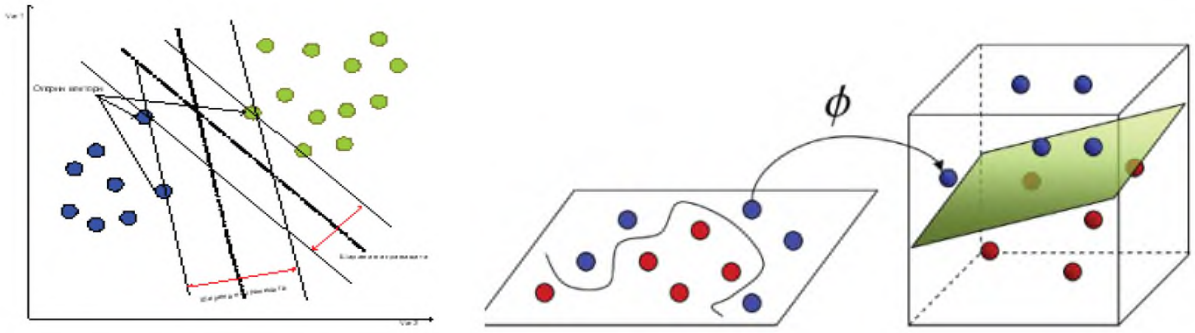


Рис. 1. Классификационная модель k-ближайших соседей

Если среди выбранных k образов из выборки количество объектов, принадлежащих к одному и тому же классу, равно или превышает S, считается, что объект принадлежит к этому классу. Если это условие выполняется для двух или более классов, неизвестный объект относится к тому классу, у которого наибольшее количество ближайших соседей. Параметр k - целое число, обычно небольшое. Количество ближайших соседей, по которым вычисляется евклидово расстояние, определяется экспериментально (рис. 1).

➢ Метод опорных векторов, или классификатор SVM (Support Vector Machines) - выполняет нелинейное преобразование исходных данных в другое более высокой размерности пространство, в котором объекты становятся линейно разделимыми (рис. 1б) [8].

➢ В методе SVM можно вычислить гиперплоскости, разделяющие классы, причём, так, чтобы расстояние между граничными гиперплоскостями двух классов было максимальным (рис. 2а).



а) Преобразование входного пространства для описания данных в методе SVM **б)**

Рис. 2. Оптимальная гиперплоскость в методе SVM: а) при линейно разделимых областях; б) при нелинейно разделимых областях.

ISSN 2181-4732
ISSN 2181-4732

В зависимости от выбранной ядерной функции Φ, могут быть построены различные типы классификаторов (линейный, RBF, полиномиальный, нейронная сеть MLP) [7, 8]. В данном случае используется ядерная функция типа RBF (Radial Basis Function) с шириной $\sigma = 0{,}625$, которая имеет следующий вид:

$$K(x, y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (3)$$

➢ **Классификационная модель «Дерево решений» - Decision Trees**

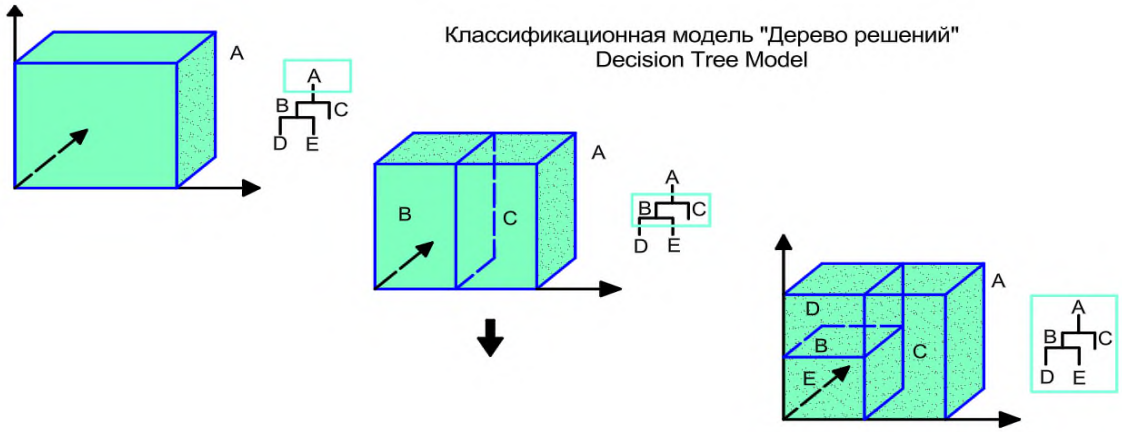


Рис. 3. Классификационная модель «Дерево решений».

Дерево решений определяет последовательность - набор логических условий if-then, на основе которых строится классификационная модель, используя статистические оценки и распределения доступных признаков. Преимущества метода Decision Trees:

- Интерпретация результатов, полученных с помощью этого метода, чрезвычайно проста (вместо линейных уравнений формируется набор условий вида *if-then*);
- Часто классификационная модель, построенная этим методом, гораздо проще, но не менее эффективна;
- Метод является непараметрическим и нелинейным;
- Не требует предварительных условий о наличии линейной зависимости между предсказываемыми и зависимыми переменными.

Для анализа пригодности отобранных комбинаций признаков метод был применён в двух вариантах классификации.

Результаты классификации и валидации

Поскольку основной целью настоящего исследования является выбор оптимального варианта классификатора из нескольких альтернатив, был применён метод кросс-валидации (Cross-validation), также называемый перекрёстной валидацией. Его задача - настройка моделей классов и последующая оценка точности, с которой созданные модели смогут применяться в конкретной задаче классификации.

В этом методе валидациіи задаётся процентное распределение образцов между двумя выборками, а имеющиеся данные классов случайным образом разделяются на набор пар: обучающая и тестовая (валидационная) выборки. Для каждой такой пары проводится обучение и тестирование классификатора, после чего вычисляется средняя общая точность по всем парам выборок - это и является итогом валидации.

Учитывая природу и характеристики кукурузных зёрен, необходимо принимать во внимание влияние стороны съёмки зерна - «гладкая» или «зародыш» - а также тот факт, что цветовые характеристики области зародыша на изображениях зерна близки к характеристикам налёта, образующегося под воздействием плесени *Fusarium Moniliforme*. С учётом этих факторов определены размеры тестовых выборок для классификации подгрупп зёрен.

ISSN 2181-4732
ISSN 2181-4732

Соответственно, при классификации групп «только гладкая» и «только зародыш» количество зёрен в отдельных подвыборках по сортам составляет: 135 - обучающая, 45 - тестовая, 150 - всего.

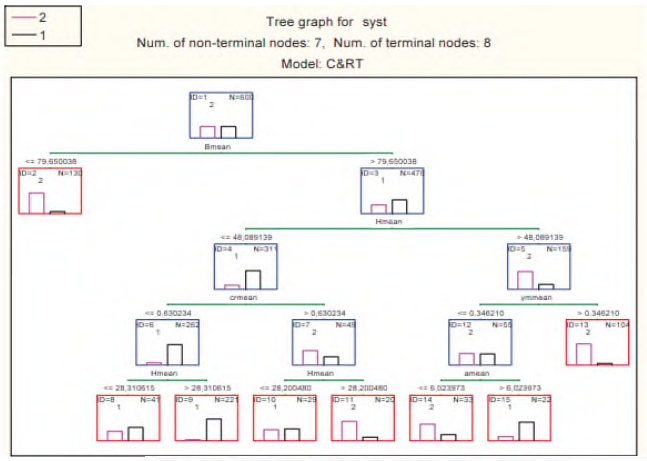
При классификации группы «общая» (включает изображения обеих сторон) - 450 - обучающая, 150 - тестовая, 600 - всего. Для сорта Русе 424: 135 - обучающая, 45 - тестовая, 180 - всего. Для сорта XM87/136: 90 - обучающая, 30 - тестовая, 120 - всего.

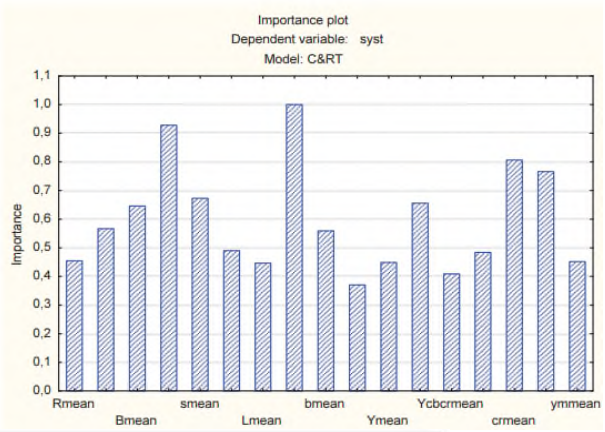
При синтезе классификаторов k-NN для каждого сорта оптимальное количество ближайших соседей k выбиралось экспериментально (рис. 1). Выбрано такое значение параметра k, при котором достигается наивысшая точность кросс-валидации, при этом тестировались 5 вариантов, соответственно для k=1,2,3,4,5.

В качестве оптимального классификатора выбирается та модель и соответствующее ей оптимальное значение параметра k, при котором достигается минимальная ошибка валидации. Можно сделать вывод, что в большинстве выборок наивысшая точность классификации достигается при методе опорных векторов, даже при меньшем количестве признаков.

На рис. 4 показано принятое решение классификатора по методу «Дерево решений», соответствующее объединённой выборке данных по сорту Кнежа 308, при этом выбран критерий хи-квадрат согласно формуле (3). На рисунке можно увидеть сгенерированные условия решения, схематическое распределение образов в выборке в соответствии с решающими условиями, а также график информативной значимости каждого признака модели.

В работе представлены обобщённые результаты классификации методом «Дерево решений» по изображениям зёрен из объединённых выборок (гладкая сторона + зародыш) для всех исследованных сортов, классифицированных методом «Дерево решений». Также проведено сравнение с точностью, достигнутой классификатором k-NN.





Tree Structure (Stat_pr_mean_kn308_all.sta)
Response: syst
Model: C&RT

	Number of nodes	Size of node	N in class 2	N in class 1	Selected category	Split variable	Criterion for child 1	Criterion for child 2	Child node 1	Child node 2
1	2	600	300	300	2	Bmean	x <= 79,6500	x > 79,6500	2	3
2		130	118	12	2					
3	2	470	182	288	1	Hmean	x <= 48,0891	x > 48,0891	4	5
4	2	311	57	254	1	crmean	x <= 0,6302	x > 0,6302	6	7
6	2	262	26	236	1	Hmean	x <= 28,3106	x > 28,3106	8	9
8		41	17	24	1					
9		221	9	212	1					
7	2	49	31	18	2	Hmean	x <= 28,2005	x > 28,2005	10	11
10		29	14	15	1					
11		20	17	3	2					
5	2	159	125	34	2	ymmean	x <= 0,3462	x > 0,3462	12	13
12	2	55	28	27	2	amean	x <= 6,0240	x > 6,0240	14	15
14		33	24	9	2					
15		22	4	18	1					
13		104	97	7	2					

Рис. 4. Решение классификатора «Дерево решений».

ISSN 2181-4732
ISSN 2181-4732

Рассчитанная точность, представляет собой процент правильно распознанных классификатором примеров относительно всех примеров и вычисляется по следующей формуле:

$$\boldsymbol{T = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%}, \quad (4)$$

где следующие обозначения:

- TP (True Positive) - действительно правильный;
- TN (True Negative) - действительно неправильный;
- FP (False Positive) - ошибочно распознанный как правильный;
- FN (False Negative) - ошибочно распознанный как неправильный.

Эти метки соответствуют классам, которые классификатор может ассоциировать с каждым входным образцом.

Заключение

В работе представлен сравнительный анализ нескольких критериев для определения диагностической значимости каждого признака — как по отдельности, так и в комбинациях. Для принятия окончательного решения относительно оптимального метода классификации должна учитываться максимальная достигнутая точность стандартными методами [9].

Установлена возможность сокращения числа признаков. На практике их количество может быть снижено - в зависимости от используемого метода распознавания:

- до 3 признаков при методе опорных векторов (SVM),
- 4–5 признаков при методе «Дерево решений»,
- 9–11 признаков при методе «k-ближайших соседей (k-NN) - при этом сохраняются сравнительно близкие значения точности классификации.

Установлено, что метод опорных векторов обеспечивает значительно более высокую скорость обработки и удовлетворительную точность при использовании меньшего количества признаков по сравнению с методом k-ближайших соседей.

Классификационная модель, построенная по методу «Дерево решений», значительно проще, но не менее эффективна. Достигнутая точность по этому методу оказалась наивысшей для шести сортов и составляет 87,5 ÷ 99,17%.

Литература

[1] Колориметрия. URL: https://www.lkmportal.com/enc/kolorimetriya.
[2] Клюев В.В., Соснин Ф.Р., Филинов В.Н. Неразрушающий контроль и диагностика : справочник. Москва: Машиностроение, 2003. 656 с.
[3] Штейнберг Т.С., Семикина Л.И., Шведова О.Г. О разработке инструментального метода оценки цвета муки, выработанной из твердой пшеницы для макаронных изделий. *Хлебопродукты*. 2014. № 1. С. 56–60.
[4] Юстова Е.Н. Цветовые измерения (Колориметрия). Санкт-Петербург: Изд-во С.-Петерб. ун-та, 2000. 397 с.
[5] Джадд Д., Вышецки Г. Цвет в науке и технике. Москва:.Мир. 1978, 592 с.
[6] Kavdir I., D.E.Guyer, Evaluation of different pattern recognition techniques for apple sorting, Biosystems Engineering 99, p.211-2129, 2008.
[7] Serkan G., O. N. Gerek et al., The search for optimal feature set in power event classification, Expert systems with applications, 2009.
[8] Pazoki A., Z. Pazoki, Classification system for rain fed wheat grain cultivars using artificial neural network,African Journal of Biotechnology Vol.10(41), p. 8031-8038, 2011.
[9] Tsang C. et.al., Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection,Pattern Recognition40,p.2373-2391, 2007.