МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского» (ННГУ)

Институт информационных технологий, математики и механики

Кафедра: Математического обеспечения и суперкомпьютерных технологий

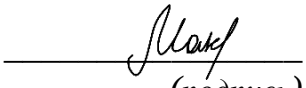
Направление подготовки: «Фундаментальная информатика и информационные технологии»
Профиль подготовки: «Компьютерная графика и моделирование живых и технических систем»

ОТЧЕТ

по учебной практике
на тему:

«Обнаружение и классификация болезней листьев растений с использованием классификатора на основе Multiclass SVM»

Выполнила:
студентка группы 381706-1
Максимова Ирина Игоревна



(подпись)

Научный руководитель:
профессор кафедры МОСТ ИИТММ,
доктор технических наук,
Турлапов Вадим Евгеньевич

_____

(подпись)

Нижний Новгород
2019

Содержание

1. Введение

Болезни растений являются одной из главных угроз глобальной продовольственной безопасности. Ежегодно болезни растений приводят к потере 10–16% мирового урожая сельскохозяйственных культур, стоимость которого оценивается в 220 миллиардов долларов [3]. Согласно прогнозу ООН, численность населения мира достигнет 9,8 миллиардов человек в 2050 году и 9,7 миллиардов в 2100 году [4]. Поэтому необходимо свести к минимуму потери мирового урожая, чтобы удовлетворить потребности в продовольствии постоянно растущего населения.

Одним из главных шагов на пути к сохранению урожая является использование эффективных методов для раннего выявления заболеваний растений.

Мониторинг сельскохозяйственных культур для выявления заболеваний растений играет ключевую роль в успешном выращивании урожая. В настоящее время основным подходом, используемым на практике, является наблюдение невооруженным глазом эксперта. Получение заключения эксперта, диагностика заболевания и обращение за консультацией к практикующим врачам - долгая, дорогая и трудоемкая практика. Кроме того, визуальная идентификация при мониторинге большого поля сельскохозяйственных культур отнимает так много времени, что порой лечить болезнь уже поздно. Поэтому в настоящее время возникает острая необходимость в создании быстрых и недорогих современных методов обнаружения заболеваний растений.

Актуальность данной работы заключается в обеспечении раннего выявления болезней листьев растений путем автоматизации процесса мониторинга сельскохозяйственный угодий. Раннее диагностирование заболевания позволит фермерам вовремя принять нужные меры по предотвращению развития и распространения болезней.

Машинное обучение представляет собой современный метод обработки изображений и анализа данных с высоким показателем точности и большим потенциалом. Машинное обучение успешно применяется в различных областях, также оно недавно вошло в область сельского хозяйства. Поэтому в данной работе в качестве классификатора болезней листьев растений будет применяться нейронная сеть.

Целью данной работы является разработка эффективного программного обеспечения для автоматического обнаружения и классификации болезней листьев растений с использованием нейросетевого классификатора.

Объектом исследования являются методы обработки изображения, текстурного анализа и машинного обучения, а **предметом** исследования – классификация болезней листьев растений по внешним симптом на ранних стадиях заболевания.

Проблематика данной работы заключается в определении качественных методов обработки изображения, текстурного анализа и машинного обучения в задаче классификации болезней растений, которые в результате дают высокий точный результат.

В рамках преследуемой цели сформулирован общий подход по созданию автоматизированной системы классификации болезней листьев растений. Предлагаемый подход состоит из четырех основных этапов:

Первый этап.

Для начала необходимо *найти базу данных*, но котором будет выполнятся обнаружение и классификация болезней листьев растений. Данные должны содержать помеченные изображения как больных, так и здоровых растений каждого исследуемого вида и сорта. Качество листовых изображений играет важную роль, и поэтому необходимо использовать надежный источник листовой базы данных. Выборка должна быть достаточно большой для корректного обучения нейронной сети.

Второй этап.

Основными возбудителями болезней культурных растений являются – паразитические бактерии, грибы и вирусы. Данные патогены вызывают отклонения жизненных процессах поражённого растения, что приводят к значительным изменениям не только его внутреннего состояния, но и его внешнего вида. Внешними признаками болезней растений являются: увядание, разрушение, мумификация и деформация органов растения, отмирание, изменение цвета тканей, появление гнили, налета, пустулы, нароста или опухоли, выделение камеди.

Поскольку появление побочных внешних эффектов сигнализирует о реакции растения на заболевание, то обнаружение подозреваемых растений может проводиться путем выявления зараженной области. В обработке изображений одним из распространенных способов выделения пораженной части листа является *сегментация*. Если подозреваемое растение действительно болеет, то пораженный участок будет определять конкретный сегмент изображения.

Третий этап.

Листья растений достаточно репрезентативны, чтобы различать виды или сорта растений и их заболевания с высокой точностью. Текстура, как тактильная, так и визуальная является главной характеристикой поверхности. Анализ текстур направлен на поиск уникального способа представления основных характеристик текстур и представления их в некоторой более простой, но уникальной форме, чтобы их можно было использовать для

надежной, точной классификации объектов. *Извлечение текстурных характеристик* — это процесс уменьшения размера данных изображения путем получения необходимой информации из сегментированного изображения. Поэтому для диагностики заболевания необходимо осуществить анализ текстуры подозреваемого листа. Наиболее распространенным вариантом получения статистических характеристик текстуры является матрица совместной встречаемости уровней серого тона (*GLCM*).

Четвертый этап.

На основе извлеченных текстурных признаков осуществляется *классификация* болезней листьев растений. В задаче классификации самым популярным подходом является использование методов машинного обучения. Наиболее популярные классификаторы — это искусственная нейронная сеть (ANN), вероятностная нейронная сеть (PNN), сверточная нейронная сеть (CNN), k-ближайший сосед (KNN) и машина опорных векторов (SVM). Каждый из них имеет свои преимущества и недостатки в задаче классификации. Поэтому необходимо провести обзор вышеперечисленных классификаторов и определить наиболее эффективный метод классификации, применяемый в исследуемой области.

Вышеизложенные этапы являются базовыми и обязательными в задаче классификации болезней листьев растений. Качественная реализация каждого этапа позволит с высокой точностью диагностировать заболевание растения.

Таким образом, с учетом предложенного подхода для достижения поставленной цели необходимо решить следующие задачи:

1. Освоить необходимый теоретический минимум. А именно:
   - Изучить статистический метод исследования текстуры на базе GLCM.
   - Ознакомиться с основными методами машинного обучения, используемыми в задачах классификации – ANN, PNN, CNN, KNN и SVM.
2. Провести обзор публикаций по теме работы. Основными темами исследуемых публикаций должны быть:
   - Сравнение нейросетевых классификаторов болезней листьев растений.
   - Методология решения задачи классификации болезни листьев растений.
3. Осуществить разбор наиболее содержательных публикаций.
4. Детально изучить каждый этап в классификации болезни листьев растений и подготовить алгоритмическую базу.
5. Найти базу данных для обучения и тестирования нейронной сети.

2. Необходимый теоретический минимум

В данном разделе будет представлен необходимый теоретический минимум, который потребуется для дальнейшего изучения исследуемой области. Необходимо освоить статистический метод исследования текстуры на базе GLCM, также ознакомиться с основными методами машинного обучения, используемыми в задачах классификации – ANN, PNN, CNN, KNN. Данные знания потребуются для дальнейшего анализа литературы по теме работы.

2.1 GLCM

Один из аспектов текстуры связан с пространственным распределением и пространственной взаимосвязью значений яркости локальной области изображения с ростом расстояния между оцениваемыми точками.

Матрица совместной встречаемости уровней серого тона *GLCM* представляет собой оценку плотности распределения вероятностей второго порядка, полученную по изображению в предположении, что плотность вероятности зависит лишь от расположения двух пикселей. Обозначим эту матрицу $P(i,j,d,\varphi)$, где i и j – яркости соседних точек на изображении, расположенных на расстоянии d друг от друга, при угловом направлении φ. Размер матрицы определяется количеством градаций яркости изображения.

Матрица GLCM обычно приводится к одному из двух видов:

1. Симметричная матрица:

$$S(i,j,d,\varphi) = P(i,j,d,\varphi) + P(i,j,d,\varphi)^T$$

2. Нормализованная матрица (матрица условной вероятности):

$$N(i,j,d,\varphi) = \frac{P(i,j,d,\varphi)}{\sum_{m,k} P(m,k,d,\varphi)}$$

По матрице GLCM вычисляется около двадцати текстурных признаков изображения.

2.2 Нейронные сети

Нейронные сети (*NN*) — это вычислительные системы со взаимосвязанными узлами, которые работают подобно нейронам человеческого мозга.

Структура NN:

1. Входной слой. Во входной слой поступают исходные данные, и передаются дальнейшим слоям. (В данной работе входной слой принимает набор пикселей изображения)
2. Конечное множество из N скрытых слоев. В каждом скрытом слое происходит взвешенное суммирование выходных сигналов предыдущего слоя и формирование выхода посредством функции активации.
3. Выходной слой. Вывод результата. В данной работе выходом будет распределения вероятностей принадлежности к определённому классу.

Сигналы между нейронами разных слоев сети передаются через соединения *синапсов*. У синапсов есть 1 параметр — *вес*. Благодаря ему, входная информация изменяется, когда передается от одного нейрона к другому.

2.2.1 Искусственная нейронная сеть

Рассмотрим *искусственные нейронные сети* (*ANN*) с точки зрения задачи классификации. В качестве образов могут выступать любые объекты: символы текста, изображения, образцы звуков и т. д. При обучении сети предлагаются различные образы с указанием того, к какому классу они относятся. Образ представляется как вектор значений признаков. При этом совокупность всех признаков должна *однозначно определять класс*, к которому относится образец. В случае, если признаков недостаточно, сеть может соотнести один и тот же образец с несколькими классами, что неверно. По окончании обучения сети ей можно предъявлять неизвестные ранее образы и получать ответ о принадлежности к определённому классу.

Топология ANN:

- Количество нейронов в выходном слое равно количеству определяемых классов. При этом устанавливается соответствие между выходом нейронной сети и классом, который он представляет.
- Когда сети предъявляется некий образ, на одном из её выходов должен появиться признак того, что образ принадлежит этому классу. В то же время на других выходах должен быть признак того, что образ данному классу не принадлежит. Если на двух или более выходах есть признак принадлежности к классу, считается, что сеть «не уверена» в своём ответе.

2.2.2 Сверточные нейронные сети

Сверточная нейронная сеть (*CNN*) — это класс глубоких нейронных сетей, который в основном используется для распознавания изображений, классификации изображений, обнаружения объектов и т.д.

Архитектура CNN:

1. Слой свертки. Этот слой выполняет точечное произведение между двумя матрицами, где одна матрица (фильтр / ядро) является набором изучаемых параметров, а другая матрица является ограниченной частью изображения.
2. Пулинг. Этот уровень предназначен исключительно для уменьшения вычислительной мощности, необходимой для обработки данных. Это делается за счет уменьшения размеров входной матрицы. В этом слое извлекаются доминирующие признаки из ограниченного количества окрестностей.
3. Полносвязная нейронная сеть. С этого момента начинается процесс классификации. После преобразования входного изображения в матрицу, содержащую основные характеристики изображения, сводим ее в один вектор-столбец. Этот вектор подается в нейронную сеть.

2.2.3 Вероятностная нейронная сеть

В *вероятностной нейронной сети* (*PNN*) операции организованы в многослойную сеть прямой связи с четырьмя уровнями:

1. Входной слой.
2. Слой шаблона. Этот слой содержит один нейрон для каждого случая в наборе обучающих данных. Слой шаблона вычисляет расстояние от входного вектора до обучающих входных векторов. Создается новый вектор, в котором элементы указывают, насколько близок полученный вход к обучающему входу.
3. Слой суммирования. Суммирует вклад для каждого класса входных данных и производит свой выход в виде вектора вероятностей.
4. Выходной слой. Выбирает максимум этих вероятностей и выдает 1 (положительная идентификация) для определенного класса и 0 (отрицательная идентификация) для нецелевых классов.

2.3 Метод опорных векторов

Основная идея *метода опорных векторов* (*SVM*) — это перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. *Разделяющей гиперплоскостью* будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

В классической версии SVM – бинарный классификатор. Но существуют различные модификации SVM, позволяющие классифицировать объекты на несколько классов.

2.4 Метод k-ближайших соседей.

Метод K-ближайших соседей (*KNN*) — метрический алгоритм для автоматической классификации объектов. Объект присваивается тому классу, который является наиболее распространённым среди k-соседей данного элемента, классы которых уже известны.

Для классификации каждого из объектов тестовой выборки необходимо последовательно выполнить следующие операции:

1. Вычислить расстояние до каждого из объектов обучающей выборки
2. Отобрать k объектов обучающей выборки, расстояние до которых минимально
3. Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей.

3. Описание базы данных

В данной работе будет использоваться набор данных взятый из открытого источника [5]. Он включает в себя 87 000 изображений здоровых и больных растений, которые подразделяются на 39 различных классов: 38 различных листовых болезней и 1 класс здоровых растений**.** В корневом каталоге находятся три версии набора данных:

- Каталог *color*: Оригинальные изображения RGB
- Каталог *grayscale*: версия градаций серого изображения
- Каталог *segmented*: RGB-изображения с сегментированным листом и цветовой коррекцией.

Каждый каталог состоит из 39 подкаталогов, с именами [*сорт растения*]_[*заболевание*] и [*сорт растения*]_[*здоровое состояние*], в которых находятся в соответствующие изображения.

Описание изображений:

- Разрешение 256x256
- Глубина цвета – 24

4. Обзор источников

Машинное обучение - самый популярный метод классификации. Часто используемые классификаторы - ANN, PNN, CNN, KNN и SVM. Поскольку у каждого из них разные преимущества и недостатки, то, для их использования нужно проанализировать соответствующую литературу и определить наиболее «эффективный» классификатор, который позволит достичь хорошей точности в определении болезни растения.

Сравнительный анализ вышеперечисленных классификаторов приводится авторами Muhammad Azfar Firdaus Azlah, Lee Suan Chua, Fakhrul Razan Rahmad, Farah Izana Abdullah и Sharifah Rafidah Wan Alwi в своей работе *«Review on Techniques for Plant Leaf Classification and Recognition»* [1].

В литературе сообщается о нескольких методах классификации заболеваний растений. Как правило, все методы включают три основных этапа: предобработка изображения, выделение признаков из изображения, классификация.

Достаточно подробно методология обнаружения и классификации болезней растений с использованием методов обработки изображений описана в статье Sandesh Raut и Amit Fulsunge *«Plant Disease Detection in Image Processing Using MATLAB»* [2].

4.1 Azlah M.A.F, Chua L.S, Rahmad F.R, Abdullah F.I, Wan Alwi S.R, “Review on Techniques for Plant Leaf Classification and Recognition”

Целью данной работы является обзор и анализ реализации и эффективности различных методов классификации растений на основе результатов предыдущих исследований в данной области. Сравниваемые классификаторы: ANN, PNN, CNN, KNN и SVM. Результатом работы является определение преимуществ и недостатков каждого рассматриваемого классификатора.

Таблица 1.

Сравнение преимуществ и недостатков классификаторов

Кл-р	Преимущества	Недостатки
ANN	• Способен различать сложные нелинейные отношения между независимыми и зависимыми переменными. • Упрощенная статистическая подготовка при выделении отличительных признаков растений.	• Склонность к переобучению • Большая вычислительная нагрузка.
CNN	• Устойчив к шуму. • Быстрый процесс распознавания	• Склонность к переобучению • Большая вычислительная нагрузка. • Нет возможности для обобщения.
PNN	• Высокая устойчивость к искажению. • Гибкость при изменении данных. • образец можно классифицировать на несколько классов.	• Длительное обучение. • Сложная схема сети. • Склонность к переобучению.
SVM	• Большой потенциал обобщения. • Высокая точность классификации. • Надежен, даже когда обучающие данные имеют некоторые искажения.	• Сложная структура алгоритма. • Медленное обучение и тестирование.
KNN	• Не требуется обучение. • Надежен в плане исследуемого пространства. • Простейший классификатор	• Чувствителен к шуму. • Ленивое обучение. • Длительный процесс тестирования.

В данном исследовании делается вывод о том, что классификатор, который игнорирует искажения, значительно улучшает технологию классификации листьев растений. В том числе делает классификацию водной фауны более качественной, поскольку водные растения могут не иметь определенной формы.

Собственные выводы. В рамках преследуемой цели, для решения задачи классификации болезней растений в своей работе я буду использовать SVM, а точнее его модификацию - MSVM. Поскольку оригинальный SVM - способен классифицировать объекты только на два класса. Этот метод классификации обладает нужными, для моей работы, качествами. Такими как высокая точность, надежность и способность обобщения, которая в случае успешного обучения позволит вернуть верный результат на основании данных, которые отсутствовали в обучающей выборке, а также неполных, зашумленных и частично искажённых данных.

4.2 Prabira K.S, Nalini K.B, Amiya K.R, “Detection & Identification of rice leaf diseases using Multiclass SVM and Particle Swarm Optimization technique”

В данной статье описан новый подход к выявлению и идентификации болезней листьев риса с помощью кластеризации K-средних, классификатора MSVM и оптимизации методом роя частиц (МРЧ). Матрица GLCM используется для выделения признаков. Классификация заболеваний выполняется с использованием классификатора MSVM, а точность обнаружения улучшается путем оптимизации данных с использованием PSO.

Предложенная методология идентификации заболевания:

1. Получение изображения
2. Предварительная обработка изображения: преобразование цветового пространства в CIE L*a*b*, повышение контрастности, изменение размера изображения.
3. Сегментация изображения с использованием алгоритма K-means clustering. Кластеризация выполняется в пространстве 'a*b*'
4. Извлечение признаков с использованием GLCM. В этом исследовании используются 13 текстурных характеристик, рассчитанных по GLCM, а именно: однородность, контрастность, среднее значение, корреляция, энергия, стандартное отклонение, асимметрия, среднеквадратичное значение (RMS), дисперсия, энтропия, гладкость, эксцесс и обратный разностный момент (IDM).
5. Оптимизация данных с использованием МРЧ. Здесь используется метод МРЧ для оптимизации больших наборов признаков.
6. Классификация болезни

С точки зрения исполнения, предложенная методология была успешно опробована и проверена на четырех видах болезней рисовых листьев. Также авторы сравнивают эффективности предложенного метода при использовании различных классификаторов: KNN, NN, SVM, МРЧ + SVM. В результате исследования полученная точность каждого классификатора составляет соответственно 77.90%, 85.64%, 90.50%, 97.91%.

Таким образом, с помощью метода PSO для оптимизации набора характеристик, SVM оказался перспективным методом для дифференциации и классификации болезней листьев риса с средней точностью 97,91%.

Собственные выводы. Благодаря разбору данной статьи мне удалось более подробно изучить подходы к реализации этапов предобработки изображения, выделения признаков и классификации. Целесообразно на этапе предобработки изображения максимально улучшить качество исследуемого изображения для наиболее эффективного извлечения информации. На

этапе выделения признаков разумно выделять как можно больше признаков. Но некоторые признаки сильно коррелируют друг с другом, например дисперсия и среднеквадратичное отклонение, и в результате не несут в себе полезной информации. Поэтому в своей работе я буду использовать меньшее число признаков. В зависимости от набора признаков будет принято решение об необходимости использовании методов оптимизации.

5. Предлагаемая методология

5.1 Предобработка изображения ботанического листа

5.1.1 Перевод RGB изображения в CIE L*a*b*

Зависимое от устройства цветовое пространство — это цветовое пространство, в котором результирующий цвет зависит от оборудования и настроек, используемых для его создания. Например, цвет, полученный с использованием пиксельных значений RGB, будет изменяться по мере изменения яркости и контрастности дисплея.

Поэтому для осуществления корректной кластеризации изображения, необходимо чтобы его цветовое пространство не являлось аппаратно-зависимым.

На сегодняшний день, одним из самых распространенных среди аппаратно-независимых цветовых пространств, обеспечивающего наилучшую сегментацию является цветовое пространство CIE L*a*b*. Поскольку оно обеспечивает быстрое время выполнения сегментации, даже в случае большого количества цветовых компонент в изображении [6].

Алгоритм преобразования RGB ↔ CIE L*a*b*:

В случае 8-битных изображений R, G и B преобразуются в формат с плавающей точкой и масштабируются для соответствия диапазону от 0 до 1.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \leftarrow \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

$$X \leftarrow X \,/\, X_n, \qquad \text{где } X_n = 0.950456$$

$$Z \leftarrow Z \,/\, Z_n, \qquad \text{где } Z_n = 1.088754$$

$$L^* \leftarrow \begin{cases} 116 * Y^{1/3} - 16, & Y > 0.008856 \\ 903.3 * Y, & Y \le 0.008856 \end{cases}$$

$$a^* \leftarrow 500\big(f(X) - f(Y)\big) + delta$$

$$b^* \leftarrow 500\big(f(Y) - f(Z)\big) + delta$$

$$f(t) = \begin{cases} t^{1/3}, & t > 0.008856 \\ 7.787t + 16/116, & t \le 0.008856 \end{cases}$$

$$delta = \begin{cases} 128\text{ , для } 8-\text{битных изображений} \\ 0\text{, для изображений с плавающей точкой} \end{cases}$$

В результате получаем $0 \le L \le 100$ и $-127 \le a, b \le 127$. Затем полученные значения дополнительно преобразуются, чтобы соответствовать диапазону от 0 до 255:

$$L \leftarrow L * 255/100,\ a \leftarrow a + 128,\ b \leftarrow b + 128$$

5.1.2 Выравнивание гистограммы

Выравнивание гистограммы - это метод обработки изображений с целью регулировки контрастности изображения путем изменения плотности распределения интенсивности гистограммы. В результате выравнивания гистограммы в большинстве случаев существенно расширяется динамический диапазон изображения, что позволяет отобразить ранее не замеченные детали. Считается, что для повышения качества цветных изображений наиболее эффективно применять процедуру выравнивания к каналу яркости [7].

Алгоритм выравнивания гистограммы:

1. Вычисление и нормализация гистограммы H.
2. Построение функции распределения H'.
3. Трансформация значений $H' * 255$ с округлением вниз.
4. Определение нового значения интенсивности пикселя $dst(x, y) = H'(src(x, y))$.

5.2 Кластеризация методом K-means

Сегментация изображения является важным и фундаментальным процессом для дальнейшего извлечения извлечения признаков. Сегментация изображения необходима, чтобы отделить область интереса от остальных частей изображения. Сегментация изображения выполняется для сегментирования пораженных и незатронутых участков листа. Метод кластеризации k-средних (k-means) — наиболее популярный метод сегментации. Алгоритму широко отдается предпочтение из-за его простоты реализации, большой скорости.

Суть кластеризации состоит в том, что все пиксели разбиваются на несколько не пересекающихся групп таким образом, чтобы объекты, попавшие в одну группу, имели сходные характеристики, в то время как у объектов из разных групп эти характеристики должны значительно отличаться. Полученные группы называются *кластерами*.

Цветовое пространство L*a*b* состоит из слоя яркости 'L*', цветового слоя 'a*', указывающего, где располагается цвет вдоль красно-зеленой оси, и цветового слоя 'b*', указывающего, где располагается цвет вдоль сине-желтой оси. Вся информация о цвете находится в слоях 'a*' и 'b*'. Поэтому достаточно выполнить кластеризацию в пространстве 'a*b*'. Исходными значениями для кластеризации являются вектора (x, y, a(x, y), b(x, y)).

Центроид — точка, которая является центром кластера.

Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

Алгоритм кластеризации методом k-средних:

1. Выбор из множества k пикселей те пиксели, которые будут центроидами соответствующих k кластеров. Выборка начальных центроидов может быть случайной.
2. Входим в цикл, который продолжается до тех пор, пока центроиды кластеров не перестанут изменять свое положение.
   2.1 Обходим каждый пиксель и смотрим, к какому центроиду какого кластера он является близлежащим.
      2.1.1 Привязываем пиксель к кластеру близлежащего центроида.
   2.2 Перебрали все пиксели. Считаем новые координаты центроидов k кластеров и проверяем координаты новых центроидов. Если они соответственно равны предыдущим центроидам — выходим из цикла

Проблематика кластеризации состоит в том, что число кластеров k должно быть заранее известно. Неправильный выбор количества кластеров сделает недействительным весь процесс кластеризации. Количество кластеров должно соответствовать данным. Определить наилучшее количество кластеров можно только эмпирическим способом, попробовав кластеризацию K-means с разным количествам кластеров k.

5.3 Выделение признаков на базе GLCM

Текстурные характеристики вычисляются на основании анализа сопряженности уровней яркости изображения для каждого выделенного сегмента, и сохраняются в виде матрицы совместной встречаемости уровней серого тона - Grey Level Cooccurence Matrix (GLCM) [8]. GLCM матрица $P(i,j,d,\varphi)$ определяется на основе соседних пар пикселей, разделенных заданным направлением φ и расстоянием d. Вычисление матриц (для каждого расстояния d и направления φ) может занимать много времени. Поэтому расстояние d и количество ориентаций φ часто ограничены небольшим количеством. Поскольку классификация мелких текстур требует использования малых значений для d, то в данной работе значение параметра d будет эквивалентно 1. Также достаточно рассматривать только четыре направления φ: 0°, 45°, 90° и 135°, поскольку противоположные направления (180°, 225°, 270° и 315°) будут учитываться в симметричной матрице.

Время вычислений может быть уменьшено путем использования уменьшенного количества уровней интенсивности, выполняемых посредством квантования изображения. Но поскольку матрицы GLCM вычисляются для отдельных кластеров изображения, в которых все пикселы схожи между собой, уровни интенсивности этих пикселей будут находится в

соответствующем узком диапазоне. Тем самым будет обеспечен небольшой размер матрицы GLCM.

Построение GLCM матрицы будет происходить для каждого канала: R, G и B. Для растений наиболее полезным оказывается красный канал. Это связано с тем, что существует функциональная связь между вегетационным индексом и красными каналом.

Как только матрицы вычислены, из них извлекаются текстурные признаки данного класса текстуры. Для этой цели Haralick, Shanmugam и Dinstein [8] предложили 14 мер. Позже Conners и Harlow [9] отметили, что только 5 из этих 14 мер являются достаточными:

1. Энергия. Измеряет текстурную однородность. Рассчитывается как:

$$\text{энергия} = \sqrt{\sum_{i,j} P_{ij}^2} \tag{1}$$

Диапазон = [0, 1]. Высокие значения энергии возникают, когда распределение уровня серого имеет постоянную или периодическую форму. GLCM менее однородного изображения будет иметь большое количество маленьких записей. Энергия равна 1 для однотонного(константного) изображения.

2. Энтропия. Эта статистика измеряет беспорядок или сложность изображения. Рассчитывается как:

$$\text{энтропия} = \sum_{i,j} P_{ij} * (-\ln P_{ij}) \tag{2}$$

Диапазон = (0, +∞]. Энтропия велика, когда изображение не является текстурно однородным, а многие элементы GLCM имеют очень маленькие значения. Сложные текстуры, как правило, имеют высокую энтропию. Также энтропия отражает количество информации в изображении: чем больше энтропия, тем больше информации в изображении, и наоборот.

3. Корреляция. Измеряет линейную зависимость уровня серого между пикселями относительно друг друга. Рассчитывается как:

$$\text{корреляция} = \sum_{i,j} P_{ij} * \frac{(i - \mu_i) * (j - \mu_j)}{\sigma_i * \sigma_j} \tag{3}$$

Где μ_i и μ_j - среднее GLCM, а σ_i и σ_j – дисперсия

$$\mu_i = \sum_{i,j} i * P_{ij}, \qquad \mu_j = \sum_{i,j} j * P_{ij} \tag{4}$$

$$\sigma_i^2 = \sum_{i,j} P_{ij} * (i - \mu_i)^2, \qquad \sigma_j^2 = \sum_{i,j} P_{ij} * (j - \mu_j)^2 \tag{5}$$

Диапазон = [-1, 1]. Корреляция принимает высокие значения, когда значения в GLCM распределены равномерно, и низкие значения в противном случае. Высокая корреляция текстуры означает высокую предсказуемость отношений пикселей. Корреляция равна NaN для однотонного изображения. Для симметричной матрицы GLCM $\mu_i \equiv \mu_j$ и $\sigma_i \equiv \sigma_j$.

4. Однородность. Возвращает значение, которое измеряет близость распределения элементов GLCM к диагонали. Рассчитывается как:

$$\text{однородность} = \sum_{i,j} \frac{P_{ij}}{1 + (i-j)^2} \tag{6}$$

Диапазон = (0, 1]. Принимает наибольшее значение, если наибольшие записи в GLCM находятся вдоль диагонали матрицы. Однородность равна 1 для диагональной GLCM.

5. Контраст. Количественно определяет локальные изменения интенсивности, присутствующие на изображении. Рассчитывается как:

$$\text{контраст} = \sum_{i,j} P_{ij} * (i-j)^2 \tag{7}$$

Диапазон = $[0\ , (size(GLCM, 1) - 1)^2]$. Контраст равен 0 для однотонного изображения. Концентрация ненулевых элементов в GLCM вокруг главной диагонали представляет собой низкоконтрасное изображение. При увеличении числа локальных вариаций интенсивностей контраст возрастает.

6. Заключение

В результате проделанной работы мне удалось набрать необходимый теоретический материал, для ввода в тему. Изучен статистический метод исследования текстуры GLCM, методы получения статистик по матрице GLCM. Получены базовые знания в области машинного обучения.

Помимо освоения теории был проведен аналитический обзор публикаций по теме работы. Подробному разбору было подвергнуто две особо содержательные публикации [1,2].

На основе полученной информации были детально изучены и описаны этапы предобработка изображения ботанического листа, перед подачей его признаков в SVM.

Также был найден набор данных для обучения и тестирования SVM.

Таким образом, на данный момент подготовлен основной алгоритмический и теоретический материал для начала практической реализации.

7. Список литературы

1. M.A.F. Azlah, L.S. Chua, F.R. Rahmad, F.I. Abdullah, S.R Wan Alwi, “Review on Techniques for Plant Leaf Classification and Recognition,” Computers 2019, 8, 77. 22 pp, doi:10.3390/computers8040077.
2. P.K. Sethy, N.K. Barpanda, A.K. Rath, “Detection & Identification of Rice Leaf Diseases using Multiclass SVM and Particle Swarm Optimization Technique,” International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-6S2, April 2019, pp 108-120.
3. S. Chakrabortya, A. C. Newton, “Climate change, plant diseases and food security: an overview,” Plant Pathology (2011), Vol. 60, Issue 1, pp 2–14, doi: 10.1111/j.1365-3059.2010.02411.x.
4. United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects 2019: Highlights (ST/ESA/SER.A/423).
5. Dataset of images. Available: https://github.com/spMohanty/PlantVillage-Dataset
6. G. Mathur, H. Purohit, “Performance Analysis of Color Image Segmentation using K-Means Clustering Algorithm in Different Color Spaces,” IOSR Journal of VLSI and Signal Processing (IOSR-JVSP), Vol. 4, Issue 6, Ver. III (Nov - Dec. 2014), pp 1-4. Available: www.iosrjournals.org.
7. G. Jeon, “Color Image Enhancement by Histogram Equalization in Heterogeneous Color Space," International Journal of Multimedia and Ubiquitous Engineering (IJMUE), Vol. 9, No. 7 (2014), pp 309-318, doi: http://dx.doi.org/10.14257/ijmue.2014.9.7.26
8. R. M. Haralick, K. Shanmugam, and I. Dinstein, ‘‘Textural features for image classification,’’ IEEE Trans. Syst., Man, Cybern., vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
9. R. W. Conners and C. A. Harlow, ‘‘A theoretical comparison of texture algorithms,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-2, no. 3, pp. 204–222, May 1980.