

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ В ЦИФРОВОЙ ФИТОСАНИТАРИИ

Ибрагимов Тагир Замилеевич, к.б.н., ведущий научный сотрудник, ФГБНУ «Всероссийский научно-исследовательский институт фитопатологии», Email: tibragimov@yandex.ru

Аннотация: для прогнозирования септориоза листьев пшеницы были использованы методы интеллектуального анализа. Разработана система, позволяющая с одним и тем же набором данных параллельно проводить прогнозирование методами искусственной нейронной сети, дерева решений и наивного байесовского классификатора. Система позволяет в интерактивном режиме настраивать расчетные параметры для каждого из методов, видеть полученные результаты и оценивать их эффективность.

Ключевые слова: цифровая фитосанитария, прогноз, септориоз листьев пшеницы, нейронная сеть, дерево решений, наивный байесовский классификатор, метод ассоциации.

Сельское хозяйство страны, как известно, переходит на новый – цифровой этап развития [1]. Современные методы интеллектуального анализа данных с целью прогноза наиболее эффективно разработаны для классификационного прогнозирования (классификации).

В области фитосанитарии это подразумевает оцифровку и последующую цифровизацию фитосанитарной диагностики, фитосанитарного мониторинга и использование компьютерных систем поддержки принятия решений.

Под системой поддержки принятия решений в защите растений в общем виде понимается интерактивная, информационно-аналитическая система, позволяющая обрабатывать и анализировать числовые массивы данных и иную информацию с помощью различных методов (включая интеллектуальный анализ и моделирование), интегрирующая модели прогноза и управления в компьютерные системы, максимально приспособленные к решению управленческих задач, и являющаяся инструментом, призванным оказать помощь лицам, принимающим решения [2].

Ключевой момент для управления защитой растений – прогнозирование фитосанитарной ситуации. Возможность предсказать развитие событий до принятия решений позволяет сделать наилучший выбор для защиты растений. Для целей прогнозирования существует достаточно большое число традиционных статистических и иных методов (различные варианты регрессий, сплайны, сценарный анализ и т. д.), а с развитием методов интеллектуального

анализа данных появились новые, например, деревья принятия решений, нейронные сети, ассоциации и ряд других [4,5].

Современные методы интеллектуального анализа данных с целью прогноза наиболее эффективно разработаны для классификационного прогнозирования (классификации). Классификация – это операция разделения различных объектов на несколько классов, которые могут быть определены правилами, границами классов или некоторой математической функцией.

Если классификация основана на взаимосвязи между известным назначением класса и характеристиками объекта, подлежащего классификации, она называется контролируемой. Если нет известных примеров класса, классификация не контролируется.

Применение интеллектуальных методов анализа информации в эпифитологии началось в 2000 г. [2]. В настоящее время существуют специализированные программные средства интеллектуального анализа данных KNIME (<http://www.knime.org>), Orange (<https://orange.biolab.si>), scikit-learn (<https://www.sklearn.org>), Weka (<http://www.cs.waikato.ac.nz/~ml/weka>), ___R (<https://www.r-project.org>) и ряде других.

Примером использования интеллектуального анализа могут служить наши разработки прогноза развития септориоза листьев пшеницы, в которых были использованы данные полевых многолетних (1994–2019 гг.) наблюдений, проведенных в отделе фитосанитарии и эпидемиологии болезней зерновых культур ВНИИ фитопатологии. На их основе создан тестовый набор данных (в формате .xlsx) для разработки различных методов интеллектуального анализа информации и сравнения их эффективности при постановке фитосанитарных прогнозов.

Для классификации фитосанитарной ситуации были использованы следующие методы – наивный байесовский классификатор, дерево решений, искусственная нейронная сеть и метод ассоциаций.

В качестве входных параметров были использованы: наблюдаемая степень развития септориоза на дату постановки прогноза (фаза развития пшеницы – 39); средняя температура и число дней с осадками на ту же дату. Выходной прогнозируемый параметр – фитосанитарная ситуация в фазе 75.

Наивный байесовский классификатор. Этот метод классификации, основанный на теореме Байеса, предполагает независимость предикторов, то есть наличие определенного признака, имеющего нормальное распределение, в классе не связано с наличием какого-либо другого признака. Например, фитосанитарная ситуация может стать эпифитотийной при определенной температуре, осадках и ряде других метеоусловий. Даже если эти факторы зависят друг от друга или от других погодных условий, все они независимо способствуют развитию болезни, и именно поэтому классификатор называется "наивным".

Наивная модель Байеса проста в построении и особенно полезна для очень больших наборов данных. Несмотря на простоту, она часто превосходит даже очень сложные методы классификации. Одно из главных преимуществ

этого подхода заключается в том, что для получения решения используется вся доступная информация и для прогноза учитывается множество атрибутов. Это делает его одним из самых полезных классификаторов для построения систем поддержки принятия решений.

В наших исследованиях для прогнозной классификации была построена система *NBCSeptOrange* с использованием возможностей пакета Orange 3.16. Точность прогноза для различных фитосанитарных ситуаций составила: депрессия – 100 %; эпифитотия – 77,1%; умеренное развитие – 77,5 %.

Дерево решений (ДР). Для задач классификации и регрессии это один из наиболее распространенных и легко понятных инструментов систем интеллектуального анализа [5].

ДР организовано в виде иерархической структуры и графически представляет собой перевернутое дерево с ветками, листьями и узлами. На ветках записаны независимые переменные (атрибуты), в листьях – значения зависимой (прогнозируемой) переменной, а в узлах – атрибуты, по которым различаются варианты. Чтобы классифицировать новую ситуацию, надо спуститься по веткам дерева до листа и получить соответствующее значение. Поэтому ДР можно использовать для представления правил в иерархической, последовательной структуре, где каждой ситуации соответствует единственное решение. При этом под правилом понимается логическая конструкция, представленная в виде "если ..., то ...". Данные могут быть любых типов.

После создания ДР можно построить прогнозные правила на естественном языке и наглядно интерпретировать полученные результаты.

С использованием программного пакета KNIME было построено ДР *DecisionTreeSept-C* и получены следующие прогнозные правила:

1. Если средняя температура >13 град. С, пораженность листьев (ф.39) >2 %, число дней с осадками > 9 , прогноз – эпифитотия;
2. Если средняя температура >13 град. С, пораженность листьев (ф.39) >2 %, число дней с осадками от 6 до 9, прогноз – умеренное развитие;
3. Если средняя температура >13 град. С, пораженность листьев (ф.39) <2 %, число дней с осадками меньше 6, прогноз – депрессия.

Полученная прогнозная классификация во многом совпадает с качественными моделями [3], но является более объективной.

Искусственная нейронная сеть. Эта математическая модель построена на принципах организации и функционирования биологических нейронных сетей и представляет собой распределенный параллельный процессор (вычислитель), состоящий из элементарных единиц обработки информации (искусственных нейронов).

Сходство работы нейронной сети с работой мозга состоит в том, что поступающие в нейронную сеть данные используются в процессе обучения, а для накопления знаний применяются связи между нейронами, называемые синаптическими весами [5]. Процедура изменения порядка синаптических весов для решения определенных задач называется алгоритмом обучения.

Системы, основанные на использовании нейронных сетей, отличаются такими свойствами как нелинейность, обучение на примерах с отсутствием предопределенной статистической модели входных данных, адаптивность (способность адаптировать свои синаптические веса к изменениям), очевидность ответа, контекстная информация [5].

В наших исследованиях для построения многослойного персептрона был использован пакет KNIME. Получена нейронная сеть, хорошо классифицирующая фитосанитарную ситуацию. Так, многослойный персептрон (4 слоя с 10 нейронами) позволяет точно спрогнозировать эпифитотийное развитие септориоза.

Правило ассоциации. Это утверждения, которые находят связь между данными в наборе. Анализ ассоциации измеряет силу совпадения между элементами. Целью этого метода интеллектуального анализа является не прогнозирование фитосанитарной ситуации в посевах зерновых культур как в классификации или регрессии, а поиск полезных правил в совпадениях элементов [5].

Правило ассоциации имеет две части: «Antecedent» (предшествующее событие, условие или правило) и «Consequent» (следствие, результат или последующее правило). Например, {средняя температура в мае больше 10 градусов C} → {развитие септориоза листьев озимой пшеницы}. Здесь «средняя температура в мае» является предшествующим событием, а «развитие септориоза листьев озимой пшеницы» – следствием. При этом необходимо отметить, что «Antecedent» – это не логическое выражение, а набор объектов из множества, с которыми связаны (ассоциированы) объекты того же множества, включенные в «Consequent» данного правила. Понятие “элемент или объект” может быть отнесено к той или иной категории представления данных (вещественное значение), поэтому концепция ассоциативного анализа может быть применена для комбинаций любых переменных.

В наших исследованиях для поиска ассоциативных правил использовался Data Mining пакет Orange 3.16.

Наиболее интересные ассоциативные связи были получены между развитием септориоза в фазе 75 и осадками 1 июня (10,2 мм), 4 июня (5,6 мм) и 7 июня (2,7 мм). Осадки менее 2 мм не вошли в ассоциативные правила.

Необходимо отметить, что этот метод достаточно трудоемкий, так как при поиске ассоциаций возникают очень большие потоки информации, которые необходимо анализировать.

Методы интеллектуального анализа фитосанитарной информации являются новыми и еще нет понимания их сравнительной эффективности. Поэтому нами с применением пакета KNIME была разработана система, позволяющая с одним и тем же набором данных параллельно проводить прогнозирование развития болезней озимой пшеницы методами искусственной нейронной сети, дерева решений и наивного байесовского классификатора. Система позволяет в интерактивном режиме настраивать расчетные параметры

для каждого из методов, видеть полученные результаты и оценивать их эффективность.

Библиографический список

1. Гордеев А. В., Патрушев Д. Н. Лебедев И. В. и др., Ведомственный проект «Цифровое сельское хозяйство»: официальное издание. – М.: ФГБНУ «Росинформагротех», 2019. – 48 с.
2. Ибрагимов Т. З., Санин С. С., // Цифровая защита растений и интеллектуальный анализ фитосанитарной информации, Защита и карантин растений, 2019, № 4, с. 15–18.
3. Санин С. С., Корнева Л. Г., Полякова Т. М., Прогноз риска развития эпифитотий септориоза листьев и колоса пшеницы // Защита и карантин растений. – 2015. – № 5. – С. 33–36
4. Gayl Creutzberg, Agriculture 3.0: A New Paradigm for Agriculture // Nuffield Canada on November 25th., 2014, p. 67
5. Vijay Kotu, Bala Deshpande / Predictive Analytics and Data Mining. Concepts and Practice with RapidMiner, – Elsevier, – 2015, – 426 pp

Data mining of digital phytosanitary

Ibragimov T. Z., Candidate of Biological Sciences, All-Russian Scientific Research Institute of Phytopathology 143050, Russia, Moscow region, Odintsovskii district, Bolshie Vyazemy, VNIIF

Abstract: *methods of data mining were used to predict the Septoria leaf blotch of wheat. A system has been developed that allows parallel forecasting with the same data set using the methods of an artificial neural network, a decision tree, and a naive Bayesian classifier. The system allows you to interactively adjust the design parameters for each of the methods, see the results obtained and evaluate their effectiveness.*

Key words: *digital phytosanitary, forecast, Septoria leaf blotch of wheat, neural network, decision tree, naive Bayesian classifier, association method.*