УДК 004.94:504.054

ДИНАМИЧЕСКИЙ ВЫЧИСЛИТЕЛЬНЫЙ ГРАФ ДЛЯ ОБОБЩЕНИЯ РАЗРОЗНЕННЫХ НАУЧНЫХ ДАННЫХ В УНИВЕРСАЛЬНУЮ МОДЕЛЬ НА ПРИМЕРЕ ПОВЕДЕНИЯ ЦЕЗИЯ В СИСТЕМЕ «ПОЧВА-РАСТЕНИЕ»

А.Н. Никитин
Заведующий лабораторией радиоэкологии ГНУ «Институт радиобиологии Национальной академии наук Беларуси», кандидат сельскохозяйственных наук

Государственно научное учреждение «Институт радиобиологии Национальной академии наук Беларуси», Республика Беларусь
E-mail: nikitinale@gmail.com

А.Н. Никитин
Заведующий лабораторией радиоэкологии Институт радиобиологии НАН Беларуси. Кандидат сельскохозяйственных наук. Проводит научные исследования в области моделирования поведения техногенных радионуклидов в естественных экосистемах.

Аннотация. Разработка эффективных методов синтеза новых научных знаний посредством создания моделей, обобщающих накопленные разнородные данные является актуальной задачей для областей с высокими затратами на получение каждого обучающего примера. В статье обосновывается использование для этих целей динамического вычислительного графа с известными уравнениями в узлах и измеряемыми или виртуальными переменными на гранях. Обучение графа осуществляется методом стохастического градиентного спуска и обратного распространения ошибки. Для вычисления ошибки используются как целевая, так и промежуточные переменные. Показано, что данный подход позволяет получить легко интерпретируемую и валидируемую модель для обучения которой используется весь набор имеющихся разнородных данных, описывающих исследуемый объект или явление.

Ключевые слова: динамический вычислительный граф, обобщающая модель, агрегация данных, радиоэкология, цезий.

Данная работа посвящена проблеме синтеза нового научного знания на основе обработки накопленных разрозненных данных. Продуктом научной деятельности являются не только новые знания, но и набор исходных и прошедших математическую и статистическую обработку данных. Сегодня многие издательства научных журналов и финансирующие исследования организации стимулируют размещение в общем доступе исходных данных, собранных при проведении исследований. С одной стороны, это увеличивает воспроизводимость исследований и позволяет верифицировать результаты независимыми учеными. С другой стороны, за счет обобщения и обработки объединенного массива данных могут быть синтезированы новые научные знания. С учетом того, что получение исходных данных нередко сопряжено с серьезными финансовыми затратами, такой подход позволяет снизить издержки на научную деятельности. А увеличение объема обрабатываемых данных повышает надежность результата и уровень обобщения. Синтезу нового научного знания с помощью обработки накопленных ранее данных препятствует отсутствие единства и унифицированности в доступных наборах. Даже при решении идентичных научных задач

набор данных (измеряемых параметров, условий эксперимента и т.п.) может различаться. А если стоит задача собрать максимально возможный объем данных, связанных с заданным объектом или явлением, то разнообразие наборов возрастает многократно, и степень пересечения между ними может различаться от полной до пустого множества.

Одним из наиболее распространенных подходов для синтеза нового научного знания из набора опубликованных результатов исследований является мета-анализ. Данный подход можно отнести, скорее, к полуколичественным. Используются в нем, как правило, конечные результаты опубликованных исследований, а не промежуточные показатели или исходные данные. Построить сколько-нибудь сложную модель исследуемого объекта с помощью методов метаанализа практически невозможно, а область ее применения лишь в незначительной мере расширяется относительно исходных публикаций. В области радиоэкологии примерами подобных подходов могут быть обобщающие публикации по коэффициентам перехода радионуклидов из почвы в растения [1;2] или метаанализ воздействия ионизирующих излучений на живые организмы [3-5].

Несколько в стороне стоит другой подход, заключающийся в объединение простых моделей в качестве подмоделей в единую, более сложную супермодель. Супермодель имеет более широкую область применения и может переносится на условия, не исследовавшиеся в некоторых из подмоделей. В области радиоэкологии примером такого подхода может быть модель Absolom [6; 7] и ее оптимизированная версия [8]. Для использования данного подхода необходимо наличие уже готовых субмоделей или наборов данных, позволяющих их построение. Основное ограничение заключается в том, что часто субмодели строятся по результатам разрозненных экспериментов, различающихся по условиям, контролируемым параметрам, объектам, и адекватность их объединения в супермодель нередко вызывает вопросы.

Таким образом, возникает потребность в разработке методического подхода, позволяющего создавать модели объектов или явлений на основе объединения разрозненных наборов данных, обладающих различной степенью полноты, различающихся по условиям и методом сбора, с максимально полным учетом каждого из фрагментов. Использование подобного подхода позволит создавать новые научные знания на основе собранных различными исследовательскими группами данных. Степень обобщения, надежность и область применения таких моделей может быть существенно шире по сравнению с агрегацией, осуществленной с использованием иных подходов.

Наша гипотеза состоит в том, что поставленная задача может быть решена с использованием динамического вычислительного графа. В узлах графа могут использоваться известные или вновь предлагаемые уравнения, описывающие связь между отдельными параметрами системы, а на ребрах – измеряемые или виртуальные переменные. Параметры уравнений в узлах вычислительного графа могут оптимизироваться с использованием стандартных методов машинного обучения с обратным распространением ошибки [9]. Но при обучении используются все возможные для данного примера ошибки: как финального в вычислительном графе значения, так и промежуточных, представленных в примерах из набора данных.

Одной из областей научных знаний, нуждающейся в эффективном подходе для обобщения накопленных научных данных в единой моделей, является радиоэкология. Данные в этой области собраны либо в регионах, подвергшихся радиоактивному загрязнению в результате аварий, либо в лабораторных условиях. Выявленные количественные закономерности лишь в ограниченной мере применимы для отличающихся природно-климатических условий. Это вызывает серьезные затруднения при организации мероприятий по ликвидации последствий инцидентов с выбросом радиоактивных веществ в окружающую среду. Универсальная модель поведения основных дозообразующих веществ в окружающей среде, формирования доз облучения человека и биоты, а также оценки эффективности

различных мер ограничения облучения позволит повысить степень готовности к реагированию на аварийные ситуации.

Обобщение разрозненных данных о поведении радионуклидов в окружающей среде в универсальной модели должно основываться на использовании известных физических, химических и биологических закономерностей, определяющих интенсивность перехода и перераспределения изотопов между компонентами экосистем. Для моделирования поведения радиоактивных веществ в системе "почва-растение", среди прочего, необходимо учитывать физико-химические формы радиоактивных выпадений, процессы изменения этих форм в почве, взаимодействие радиоактивных изотопов с ионами, органическими молекулами и кристаллическими структурами в почве, влияние на эти процессы микроорганизмов и корневых выделений растений, механизмы корневого поглощения элементов, физиологически и молекулярно-биохимические процессы в растениях. Учет и анализ всех этих факторов в рамках одного эксперимента практически невозможен. Но объединение результатов разнообразных экспериментов, посвященных данным проблемам, имеет существенное фундаментальное и прикладное значение.

Для проверки выдвинутой гипотезы нами выбрана очень упрощенная (минимально жизнеспособная) модель поведения 137Cs в системе "почва-растение" (рис. 1). Анализ доступных наборов данных по данному предмету указывает на то, что в них могут присутствовать как все входные и выходные переменные (очень редко), так и лишь отдельные элементы в различных сочетаниях.

Многие классические работы, посвященные проблеме накопления радиоактивных изотопов цезия растениями содержат данные о валовом содержании в почве 137Cs или 134Cs и обменного калия. Ряд авторов не уделял достаточное внимание роли K в корневом поглощении цезия, поэтому они обнародовали данные только по содержанию радиоактивного изотопа в почве и растениях. Значительное количество работ посвящено химическим процессам, влияющим на переход цезия в растворимую форму, в них мы можем найти данные по соотношению различных форм нахождения 137Cs в почве и факторам, влияющим на него. Среди рассматриваемых переменных наиболее трудоемкой в определении является содержание 137Cs в почвенном растворе, поэтому можно найти публикации, включающий практически полный набор данных, но без этой переменной. Кроме того, проведено достаточно большое количество лабораторных экспериментов с гидропонной культурой, посвященных исследованию интенсивности поступления в растения радиоактивных изотопов цезия из раствора и влиянию на нее содержания в растворе K+ и других ионов.

Можно использовать несколько различных способов включения этих данных в единую модель. Мы сразу отбрасываем вариант с фильтрацией только тех примеров, которые содержат полный набор переменных. Это сильно ограничивает набор данных для обучения модели и, по сути, не привносит ничего нового.

Второй способ заключается в фильтрации данных, содержащих полный набор входных и выходных переменных, значимых с практической точки зрения (валовое содержание 137Cs в почве, концентрация обменного калия и удельная активность 137Cs в растении, в нашем случае) и обучении модели только на них. Поскольку связь между входными и выходными переменными может быть довольно сложной, для построения модели целесообразно использовать искусственную нейронную сеть, метод опорных векторов или модели, основанные на композициях решающих деревьев. Недостатком такого способа является отказ от определенной доли данных, довольно существенной в большинстве случаев, при одновременной потребности в большом их объеме, поскольку подобные модели имеют значительное количество обучаемых параметров.

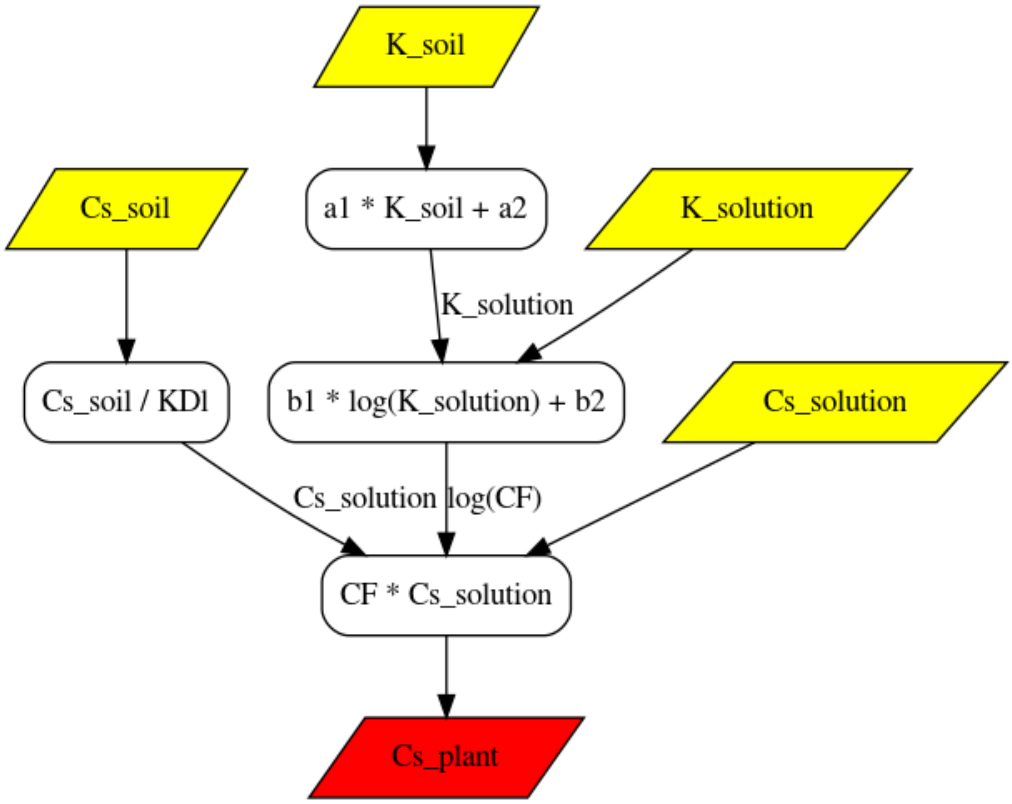

Рисунок *1.* – Упрощенная модель поведения цезия в системе "почва-растение". Условные обозначения: K_{soil} – содержание обменного калия в почве, $K_{solution}$ – концентрация калия в почвенном растворе, Cs_{soil} – удельная активность ^{137}Cs в почве, $Cs_{solution}$ – удельная активность ^{137}Cs в почвенном растворе, Cs_{plant} – удельная активность ^{137}Cs в растении, CF (коэффициент концентрирования ^{137}Cs), KDl (коэффициент распределения ^{137}Cs между твердой фазой и почвенным раствором), a1, a2, b1, b2 – обучаемые параметры модели. Желтые параллелограммы – входные переменные, Красный параллелограмм – выходная переменная.

В настоящее время наиболее распространенным подходом является обучение субмоделей для отдельных звеньев модели на доступных для этого поднаборах данных. Можно также обучать объединения из нескольких субмоделей в случае пропусков в промежуточных переменных. При объединении в модель необходимо провести усреднение обученных параметров с весами, пропорциональными надежности используемых данных. Недостатком этого способа является потенциальная несовместимость субмоделей, обученных на наборах данных, собранных в сильно отличающихся условиях. Кроме того, могут возникнуть трудности в обучении и последующем объединении композиций субмоделей, обученных на различных наборах данных, с различными вариантами пропуска промежуточных переменных.

Обучение динамического вычислительного графа на полном наборе собранных данных является способом, практически лишенным перечисленных выше недостатков. При работе с хорошо изученной проблемой в узлах вычислительного графа могут использоваться известные функции, связывающие отдельные переменные, а для инициализации – параметры этих функций, оцененные с той или иной степенью достоверности. Эти знания являются надежной отправной точкой для обучения модели.

Поскольку реальные численные значения переменных, используемых в модели, могут отличаться на порядки и такой же диапазон варьирования у параметров уравнений, обучение модели может быть крайне неэффективным. Для преодоления этой проблемы следует

использовать стандартный подход машинного обучения – стандартизация значений. Но прибегать к нормализации ($\mu = 0$, $\sigma = 1$) в данном случае неприемлемо, поскольку многие переменные, описывающие реальные состояния и качества, никогда не принимают отрицательные решения и в используемых уравнениях могут использоваться такие функции, как логарифм или квадратный корень. Оптимальным решением в данном случае может быть деление значение на максимальное или среднее в ряду данных. При выборе такого решения легко переводить параметры уравнений в стандартизированный вид и обратно.

Библиотека PyTorch является удобным средством для решения поставленной задачи. Динамический вычислительный граф в ней может быть сформирован в несколько строчек кода. Он легко подстраивается под имеющийся набор данных и имеет встроенные функции вычисления градиентов и обучения методом обратного распространения ошибки. Имеется возможность сложения нескольких ошибок, что немаловажно в данном случае, поскольку перед нами стоит задача использовать для обучения как выходные, так и промежуточные переменные.

Для апробации данного подхода в PyTorch был построен динамический вычислительный граф с архитектурой, представленной на рисунке 1, с возможностью подстраиваться под каждый обучающий пример в зависимости от входящих в его состав переменных и сложением ошибок как от выходной переменной (удельная активность 137Cs в растении), так и от промежуточных (содержание K и Cs в почвенном растворе). При этом ошибкам от промежуточных значений придавались в пять раз более низкие веса по сравнению с финальными, в соответствии с практической важностью данных переменных. В качестве функции ошибки использовано среднеквадратическое отклонение. Метод обучения – стохастический градиентный спуск; методы основанные на одновременном обучении на всем наборе данных или подвыборках здесь неприменимы, поскольку граф должен подстраиваться под каждый пример.

В испытаниях метода использованы искусственно сгенерированные реалистичные данные, полученные с помощью модели [8] и добавлением шума различной интенсивности к каждой переменной и параметру. Данные были стандартизированы делением на максимальное значение. После этого случайным образом удалено 5, 10 и 20% значений каждой переменной. Набор данных включал 1000 примеров, которые перед обучением разделялись на обучающую и проверочную выборки в соотношении 4:1.

Значения параметров модели были инициализированы единицами. Следует отметить, что предлагаемая модель практически не нуждается в регуляризации, поскольку учет ошибок от промежуточных переменных позволяет удерживать значения параметров в приемлемом диапазоне.

Предварительный анализ показал, что для обучения достаточно 500–1000 эпох. Среднеквадратичная ошибка модели на валидационной выборке при обучении на полном наборе данных, при удалении 5, 10 и 20% случайных значений каждой переменной оказалась практически идентичной – 0,012–0,013. Можно лишь отметить, что увеличение степени разрозненности данных в выборке несколько снижает скорость обучения модели (рисунки 2 и 3). Полученный результат свидетельствует о высокой эффективности предложенного подхода для обобщения разнородных данных в единую вычислительную модель.

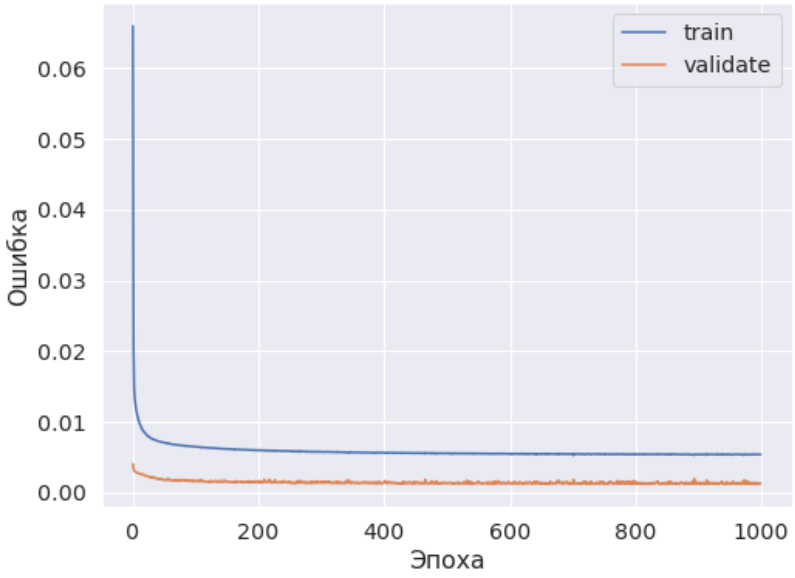

Рисунок *2.* – Динамика обучения модели на данных без пропусков. Примечание: ошибка на обучающей выборке является взвешенной суммой среднеквадратичных ошибок по выходной и промежуточным переменным, ошибка на валидационной выборке представляет среднеквадратичную ошибку только на выходной переменной

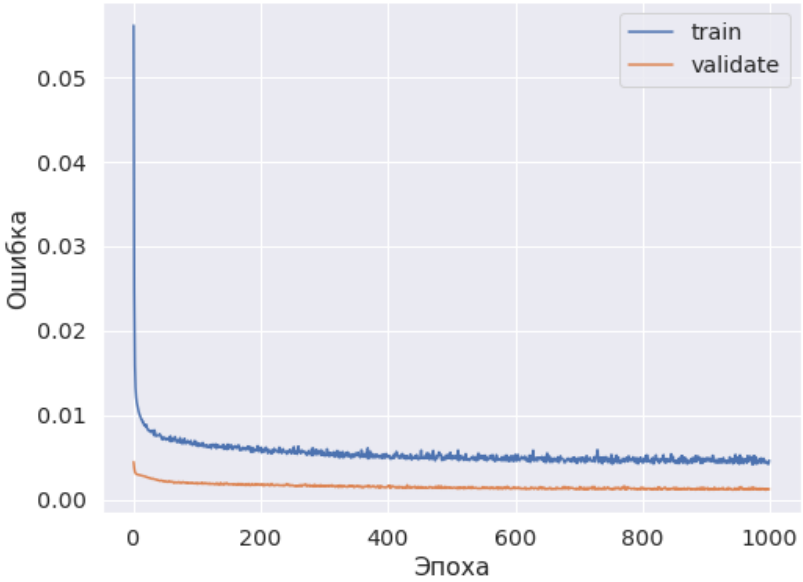

Рисунок 3. – Динамика обучения модели на данных при удалении 20% случайных значений в каждой из переменных. Примечание: ошибка на обучающей выборке является взвешенной суммой среднеквадратичных ошибок по выходной и промежуточным переменным, ошибка на валидационной выборке представляет среднеквадратичную ошибку только на выходной переменной

Таким образом, предложен подход, позволяющий создавать количественную модель, обучаемую на наборе разрозненных данных. Данный подход актуален для решения научных и прикладных задач, где получение обучающих примеров сопряжено с существенными финансовыми или иными издержками, что заставляет в максимальной степени использовать

накопленные ранее данные. Причем получение этих данных могло быть сопряжено с решением отличающихся друг от друга задач. Следует также сказать, что подобные модели используют ограниченное количество обучаемых параметров, что накладывает менее жесткие нижние ограничения на объем данных.

Использование динамического вычислительного графа позволяет обучать модель на всех данных, практически без исключения. При этом для обучения используются не только все доступные примеры, но и промежуточные значения в примерах. Ошибкам по наиболее важным с практической или иной точки зрения переменным придаются более высокие веса при обучении. Обученная модель обладает высокой степенью согласованности составных частей, а область ее определения и степень обобщения шире, чем у каждого из наборов данных в отдельности. В области радиоэкологии подобные модели могут использоваться при загрязнении радионуклидами территорий с разнообразными природно-климатическими условиями, а также применяться для прогноза поведения радионуклидов на фоне изменяющихся погодно-климатических условий.

Облегчает обучение описанного динамического вычислительного графа возможность использования известных параметров уравнений в качестве отправной точки при обучении. Состояние близкое к минимуму ошибки повышает вероятность нахождения глобального минимума и снижает риск переобучения. После обучения полученные параметры можно сравнить с опубликованными значениями и сделать дополнительное заключение об адекватности модели.

В отличие от многих других методов машинного обучения, предлагаемая модель полностью интерпретируема. Она может использоваться как целиком, так и отдельными фрагментами, давая на выходе значения, соответствующие переменным, характеризующим реальные явления или предметы.

Созданная модель может включаться в качестве составной части в более крупную, приобретая возможность обучаться на еще более широком наборе данных. В частности, модель поведения радионуклидов в системе "почва-растение" может включаться в модель миграции радиоактивных изотопов по пищевым цепям, формирования доз облучения населения и влияния инкорпорированных радионуклидов на здоровье. Подобные инструменты востребованы при проектировании аварийных ситуаций и защитных мер.

Мы полагаем, что данная модель содержит в себе дополнительную информацию, которая может быть использована для ее дальнейшего совершенствования. Ошибки по промежуточным переменным на валидационной выборке указывают на слабые узлы вычислительного графа, в которых можно попытаться изменить уравнения для получения более надежных результатов. Сумма квадратов градиентов на обучаемых параметрах указывает на надежность их оценки на использованном наборе валидационных данных. Внедрение этих приемов является ближайшей задачей развития предложенного подхода.

Список литературы

[1.] IAEA. Quantification of radionuclidetransfer in terrestrialand freshwater environmentsfor radiological assessments / IAEA. – Vienna: International atomic energy agency, 2009. – 616 p.

[2.] IAEA. Handbook of parameter values for the prediction of radionuclide transfer in terrestrial and freshwater environments. Technical report No. 472. / IAEA. – Vienna: International atomic energy agency, 2010. – 197 p.

[3.] Sazykina T.G. Non-parametric estimation of thresholds for radiation effects in vertebrate species under chronic low-let exposures / T.G. Sazykina, A.I. Kryshev, K.D. Sanina // Radiation and Environmental Biophysics. – 2009. – Vol. 48. – № 4. – P. 391-404.

[4.] Fesenko S. Comparative radiation impact on biota and man in the area affected by the accident at the chernobyl nuclear power plant / S. Fesenko [et al.] // Journal of Environmental Radioactivity. – 2005. – Vol. 80. – № 1. – P. 1-25.

[5.] Real A. Effects of ionising radiation exposure on plants, fish and mammals: Relevant data for environmental radiation protection / A. Real [et al.] // Journal of Radiological Protection. – 2004. – Vol. 24. – № 4A. – P. A123-A137.

[6.] Absalom J.P. Predicting soil to plant transfer of radiocesium using soil characteristics / J.P. Absalom [et al.] // Environmental Science & Technology. – 1999. – Vol. 33. – № 8. – P. 1218-1223.

[7.] Absalom J. Predicting the transfer of radiocaesium from organic soils to plants using soil characteristics / J. Absalom [et al.] // Journal of Environmental Radioactivity. – 2001. – Vol. 52. – № 1. – P. 31-43.

[8.] Tarsitano D. Evaluating and reducing a model of radiocaesium soil-plant uptake / D. Tarsitano, S. Young, N. Crout // Journal of Environmental Radioactivity. – 2011. – Vol. 102. – № 3. – P. 262-269.

[9.] Rojas R. The backpropagation algorithm / R. Rojas // Neural networks: A systematic introduction : vols. – Berlin, Heidelberg: Springer Berlin Heidelberg, 1996. – P. 149-182.

DYNAMIC COMPUTATIONAL GRAPH FOR GENERALIZATION BITTY SCIENTIFIC DATA IN UNIVERSAL MODEL USING BEHAVIOR OF CESIUM IN «SOIL-PLANT» SYSTEM AS EXAMPLE

A. N. Nikitin
Head of the Laboratory of Radioecology in the Institute of Radiobiology of the National Academy of Sciences of Belarus

State Scientific Institution «Institute of Radiobiology of the National Academy of Sciences of Belarus», Republic of Belarus
E-mail: nikitinale@gmail.com

Abstract. The development of effective methods for synthesizing new scientific knowledge by creating models that summarize the accumulated heterogeneous data is an urgent task for areas with high costs for obtaining each training example. The article suggest dynamic computational graph with known equations in nodes and measured or virtual variables on the edges for this purpose. The computational graph is trained by the method of stochastic gradient descent and back-propagation of the error. The error is calculated both on the target and intermediate variables. The article prove that proposed approach allows to obtain an easily interpreted and validated model for training which uses the entire set of available heterogeneous data describing the object or phenomenon being studied.

Keywords: dynamic computation graph, generalized model, data aggregation, radioecology, cesium.