УДК 004.89

Н.М. Лыченко, *nlychenko@mail.ru*
А.В. Сороковая *nastusha24sh-g@yandex.com*
Институт машиноведения и автоматики НАН КР, Бишкек, Кыргызстан

ПРОГНОЗИРОВАНИЕ КЛАССОВ ИНДЕКСА КАЧЕСТВА ВОЗДУХА Г. БИШКЕКА С УЧЕТОМ НОВЫХ ДАННЫХ 2020 – 2021 ГГ. НА БАЗЕ LSTM-НЕЙРОСЕТЕВОГО КЛАССИФИКАТОРА

Разработаны прогнозные модели индекса качества воздуха AQI г. Бишкека на основе классификаторов на базе LSTM-сети. Модели позволяют в зависимости от метеорологических условий и предшествующей истории значений AQI прогнозировать класс AQI из возможных четырех интегрированных классов: AQI≤50/ 50<AQI≤100 / 100<AQI≤150/ AQI>150. Прогноз возможен до четырех суток вперед с точностью не менее 77%.

Ключевые слова: классификация, индекс качества воздуха, LSTM-нейронная сеть, точность прогноза

Введение. Основываясь на данных об уровне загрязненности атмосферного воздуха города Бишкека, публикующихся на сайте «AirNow» [1] с февраля 2019 г., были разработаны различные модели прогноза уровня загрязненности воздуха, основные особенности и возможности которых представлены в [2]. Одна из этих моделей – модель среднесрочного прогноза класса индекса качества воздуха (Air Quality Index, AQI [3]) по метеорологическим данным и предшествующей истории значений AQI представлена в [4]. В этой работе задача прогноза AQI была сформулирована как задача прогноза класса AQI в соответствии с принятой международной классификацией [3], и были показаны возможности прогнозирования различных нейросетевых классификаторов, таких, как многослойный перцептрон, обычная рекуррентная сеть и LSTM-сеть (Long Short-Term Memory). Наилучшую точность в прогнозировании класса AQI показала LSTM-сеть. Объем располагаемых авторами данных позволил произвести классификацию AQI лишь по двум интегрированным классам, условно названных: «Хороший» (AQI <= 100) и «Нездоровый» (AQI >100). Представленные в [4] вычислительные эксперименты с LSTM-сетью показали, что «лучшую точность прогнозирования дал классификатор, учитывающий историю данных глубиной 12–16 шагов (1,5 – 2 дня), при этом прогноз AQI возможен до 4 дней вперед с точностью 88–90%» [2].

За время, прошедшее после написания работы [4], накопились новые данные, позволившие увеличить объем обучающей выборки, что в свою очередь дало возможность поставить задачу классификации AQI на большее число классов.

В настоящей работе изложены результаты прогнозирования AQI г. Бишкека на основе построенных классификаторов на базе LSTM-сети, позволяющих производить прогноз класса AQI из возможных трех и четырех интегрированных классов, а также обладающих улучшенными обучающими свойствами.

Модель LSTM-классификатора. В работе [4] для решения задачи прогнозирования была выбрана LSTM-сеть, которая учитывает исторические данные и оценивает их в зависимости от временного удаления вектора входа до прогнозируемого выхода [5].

Структура нейросетевого классификатора аналогична LSTM-сети, выбранной в [4]:

1. Входной слой, который принимает последовательность векторов признаков длиной *S.*
2. Первый скрытый слой – слой прямого распространения с числом нейронов 100 и тангенциальной функцией активации, отображает входные векторы в векторы большей длины, для дальнейшего внесения шума в данные.
3. Второй скрытый слой – слой регуляризации, который меняет некоторый процент значений выхода предыдущего слоя для предотвращения переобучения (вносит шумы).
4. Третий скрытый слой – LSTM-сеть с 50-нейронными модулями и тангенциальной функцией активации как основной классифицирующий слой.
5. Четвертый скрытый слой – слой прямого распространения с числом нейронов 10 и тангенциальной функцией активации.
6. Выходной слой – слой с количеством нейронов, соответствующим числу прогнозируемых классов и функцией активации *SOFTMAX*. Функция взвешивает входы и предсказывает вероятности активации каждого нейрона. При этом сумма выходов нейронов всегда равна 1.

Все слои сети полносвязные, т.е. каждый нейрон имеет связь с каждым предыдущим нейроном, а для рекуррентных слоев (LSTM-сеть) каждый вход слоя также связан с каждым выходом слоя [4].

Входной вектор классификатора включает в себя по меньшей мере 6 параметров – температуру воздуха, температуру точки росы, атмосферное давление на уровне станции, относительную влажность, скорость ветра, значение AQI. Входные данные обрабатываются и нормализуются перед подачей на вход нейронной сети.

Выходной вектор, к которому должен приближаться выход классификатора, определяется параметрами, характеризующими вероятность отнесения выхода классификатора к соответствующему классу.

Для остановки обучения классификатора используется порог изменения функции потерь: если на протяжении 20 эпох функция потерь меняется меньше чем на 0,01, обучение прекращается [6].

При проведении вычислительных экспериментов по прогнозированию класса AQI варьировались следующие параметры:

- S – длина последовательности векторов исторических данных входных векторов классификатора;
- P – глубина прогноза (на сколько шагов вперед прогнозируется AQI). Шаг прогноза – 3 часа.

Для каждого эксперимента подготовлены 2 файла с историческими наблюдениями AQI [1] и историческими наблюдениями погодных условий (температура воздуха, температура точки росы, атмосферное давление на уровне станции, относительная влажность и скорость ветра [7]). Программная система считывает данные, соотнося их по времени, нормализует данные по заданной функции нормализации, по заданным S и P генерирует все возможные последовательности входных векторов длиной S и сопоставляет им значения AQI в моменты времени, удаленные на P шагов от момента последних значений этих последовательностей. Группы входных признаков сбалансированы: из выборки примеров каждого класса случайным образом выбрано такое же количество примеров, которое содержится в наименьшей выборке.

Для оценки точности классификатора (по сути точности прогнозирования), так же, как в [4], формируются матрицы ошибок, а затем вычисляются соответствующие метрики (1–3) [8]. Матрица ошибок A – матрица размера $n * n$, где n – количество классов, представленных в выборке. Элемент матрицы A_{ij} содержит значение, показывающее, сколько раз классификатор определил класс *j* как класс *i*.

На основе матрицы ошибок вычисляются:

- точность в пределах класса – доля объектов, действительно принадлежащих классу относительно числа объектов, которые классификатор определил к этому классу:

$$precision_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{c,i}} \quad (1)$$

- полнота в пределах класса – доля объектов, определенных классификатором к классу, относительно всех документов, действительно принадлежащих этому классу:

$$recall_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{i,c}} \quad (2)$$

Затем вычисляется метрика, объединяющая точность и полноту, так называемая F-мера, которая и характеризует точность классификации (точность прогнозирования):

$$F = \frac{2*precision*recall}{precision+rec} \quad (3)$$

Анализ и нормализация данных. Для улучшения обучающих свойств сети были проанализированы метеорологические признаки (факторы) с 06.02.2019 до 31.03. 2020 года. Как оказалось, не все признаки распределены нормально. На рисунках 1-2 показаны распределения значений температуры и температуры точки росы в указанный период.

Как видно, распределения температуры и точки росы имеют по два явных пика, что указывает на биномиальное распределение, характеризующееся суперпозицией двух нормальных распределений. Предположительно эти распределения соответствуют двум сезонам – холодному (с ноября по март) и теплому (с апреля по октябрь). Исходя из этого предположения, была проведена декомпозиция данных температуры и температуры точки росы по сезонам (рисунки 3,4,5,6). При визуальной оценке графиков можно считать, что указанные параметры распределены нормально внутри одного сезона. Это позволит

применить Z-норму для нормализации данных при формировании соответствующих входных векторов классификатора.

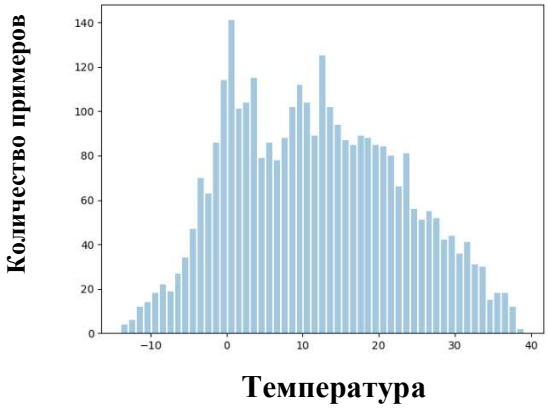


Рис. 1. Распределение значений температуры

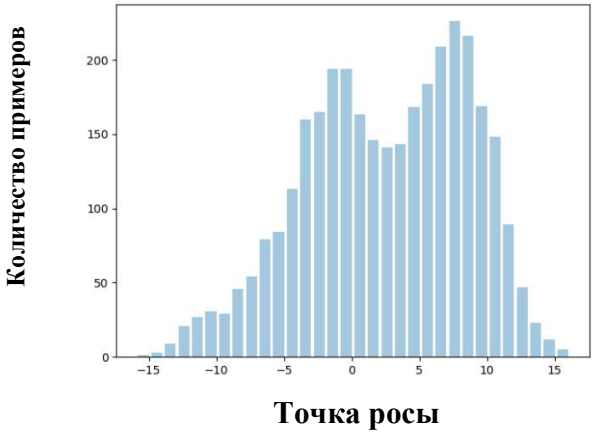


Рис. 2. Распределение значений температуры точки росы

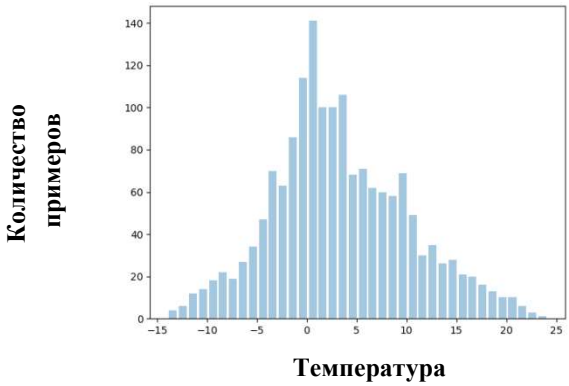


Рис. 3. Распределение значений температуры в холодный сезон года

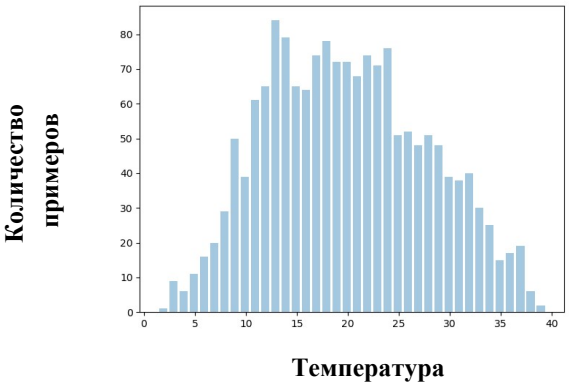
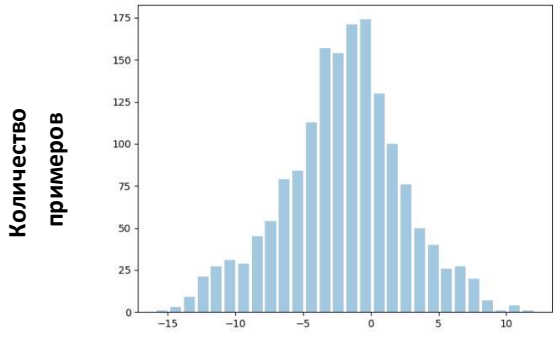


Рис. 4. Распределение значений температуры в теплый сезон года



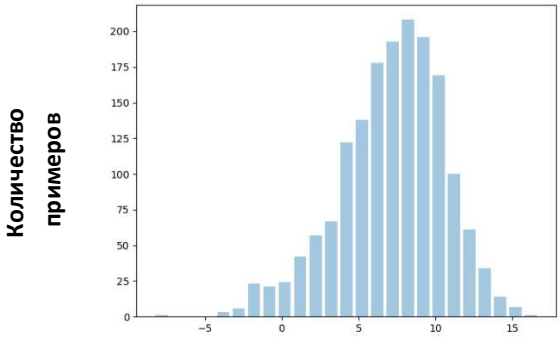


Рис. 5. Распределение значений температуры точки росы в холодный сезон года

Рис.6. Распределение значений температуры точки росы в холодный сезон года

Распределение значений атмосферного давления на уровне станции весьма близко к нормальному. Этот признак также нормализуется с помощью Z-нормы. Для относительной влажности, которая принимает значения от 0 до 100 % (от 0 до 1 – в относительных единицах), нормализации не требуется. Скорость ветра принимает всего несколько дискретных значений, нормализовать их не имеет смысла.

Классификация AQI на три интегрированных класса. В связи с введением ЧП и карантина на территории г. Бишкека в период с конца марта 2020 года и напряженной обстановкой до августа 2020 года нагрузка на общественный транспорт была снижена, воздух был чище среднего в этот период года, что могло внести существенный шум в данные, используемые для обучения. Было решено не использовать для обучения данные за период с 1 апреля 2020 года до 30 июля 2020 года. Таким образом, классификация AQI на 3 интегрированных класса была выполнена для объединенного периода наблюдений, включающего периоды с 06/02/19 по 31/03/20 и с 01/08/20 по 22/11/20.

При анализе данных наблюдений AQI в указанный период было подсчитано количество наблюдений, указывающих на определенный класс AQI и их процентное соотношение (таблица 1 и рисунок 7(а)). Из анализа количества наблюдений следует,

Таблица 1. Распределение AQI по классам для объединенного периода наблюдений 06/02/19–31/03/20 + 01/08/20–22/11/20.

Класс AQI	Количество наблюдений	Относительная доля
Хороший	695	0,165
Умеренный	2406	0,572
Нездоровый для чувствительных групп	557	0,132
Нездоровый	456	0,108
Очень нездоровый	68	0,016
Опасный	23	0,005

что в подавляющем числе зафиксированы значения AQI класса “Умеренный”. Количества наблюдений для классов “Хороший”, “Нездоровый для чувствительных групп”, “Нездоровый”, “Очень нездоровый”, “Опасный” недостаточно для решения задачи классификации AQI по всем классам. Однако можно рассмотреть задачу классификации AQI на 3 объединенных класса: «Умеренный» (объединяет «Хороший» и «Умеренный»), «Нездоровый для чувствительных групп» и «Опасный» (объединяет «Нездоровый», «Очень нездоровый» и «Опасный») (рисунок 7(б)). При таком распределении данных, если всегда

предсказывать “Умеренный” класс, то точность такого «наивного» прогноза составит примерно 74%. Значит, классификатор, прогнозирующий с точностью >74%, можно считать эффективным.

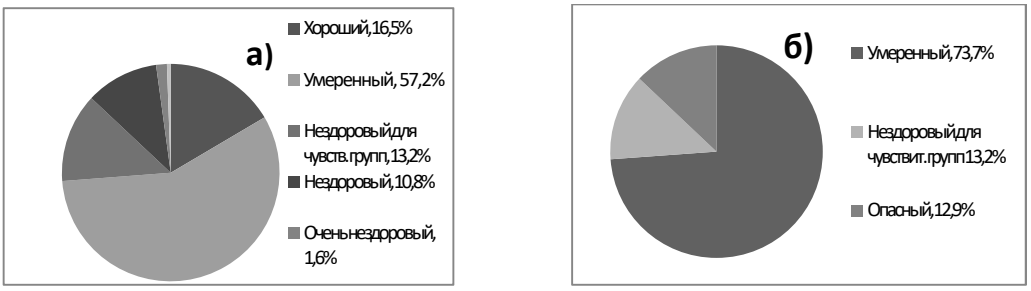


Рис. 7. Распределение наблюдений по классам AQI (а) и распределение наблюдений по трем интегрированным классам AQI (б) для объединенного периода 06/02/2019– 31/03/2020 и 01/08/2020 – 22/11/2020

За основу классификатора был взят представленный в [4] классификатор, в который был добавлен еще один выход, соответствующий добавленному классу. Таким образом, каждый выход указывает вероятностный прогноз для классов “Хороший”, “Нездоровый для чувствительных групп” и “Опасный” соответственно.

Для повышения точности прогноза и эффективности обучения сети учтена сезонность: каждый из входов температуры и точки росы расщепляется на 2 входа – в холодный период активируется один вход и на второй подается ноль, в теплый период – наоборот. При этом, поскольку границы сезонов примерно соответствуют моментам включения и выключения центрального отопления в городе Бишкека, за холодный период принят отопительный сезон, за теплый – неотопительный.

В таблице 2 представлены результаты классификации на 3 интегрированных класса. Из экспериментов видно, что только классификаторы, учитывающие историю данных глубиной 16 шагов (двое суток), дали точность прогноза более 74%.

Также в таблице 2 представлены результаты классификации для двух классов в сравнении с результатами, полученными ранее [4], а также с учетом сезонов и без их учета. Как следует из анализа таблицы, точность классификации на два класса с добавлением фактора «сезонность» несколько повысилась.

Матрица ошибок для одного из экспериментов S=16, P=8 представлена в таблице 3. Из таблицы 3 видно, например, что модель классифицировала как «Умеренный» класс: 83 случая AQI, в действительности принадлежащих «Умеренному» классу, 15 случаев AQI, принадлежащих классу «Нездоровый для чувствительных групп», 2 случая AQI, принадлежащих «Опасному» классу.

Таблица 2. F-мера точности моделей классификации на 2 и 3 класса в зависимости от длины последовательности входных векторов (*S*) и глубины прогноза (*P*).

S	**P**	**2 класса F-мера 06/02/2019-31/03/2020 (без сезонов) [4]**	**2 класса F-мера 06/02/2019-31/03/2020 (с сезонами)**	**2 класса F-мера 06/02/2019-31/03/2020 01/08/2020-31/10/2020 (с сезонами)**	**3 класса F-мера 06/02/2019-31/03/2020 01/08/2020-31/10/2020 (с сезонами)**
2	2	0.8842	-	-	-
2	4	0.8696	0.8842	0.8823	0.6558
2	8	0.8617	0.8752	0.8715	0.6307
2	16	0.8408	0.8589	0.8530	0.5983
2	24	0.8615	0.8655	0.8636	0.6136
2	32	0.8387	0.8480	0.8511	0.6115
4	2	0.8748	-	-	-
4	4	0.8787	0.8953	0.8897	0.6939
4	8	0.8753	0.8890	0.8942	0.6620
4	16	0.8684	0.8828	0.8808	0.6521
4	24	0.8585	0.8899	0.8943	0.6435
4	32	0.8552	0.8816	0.8739	0.6466
8	2	0.8976	-	-	-
8	4	0.8924	0.9272	0.9147	0.7196
8	8	0.9011	0.9234	0.9206	0.7212
8	16	0.8955	0.9284	0.9303	0.6964
8	24	0.8889	0.9318	0.9265	0.7152
8	32	0.8824	0.9251	0.9163	0.6948
12	2	0.9018	-	-	-
12	4	0.9021	0.9317	0.9281	0.7302
12	8	0.9093	0.9284	0.9324	0.7176
12	16	0.8972	0.9261	0.9316	0.7239
12	24	0.8958	0.9331	0.9367	0.7387
12	32	0.8998	0.9375	0.9343	0.7553
16	2	0.9120	-	-	-
16	4	0.8885	0.9385	0.9320	0.7444
16	8	0.9085	0.9323	0.9389	0.7605
16	16	0.9051	0.9335	0.9344	0.7407
16	24	0.9083	0.9356	0.9396	0.7538

16	32	0.9115	0.9270	0.9332	0.7493

Таблица 3. Матрица ошибок модели с параметрами S=16, P=8.

	Ожидаемый класс			
Спрогнозированный класс		Умеренный	Нездоровый для чувств.	Опасный
	Умеренный	83	15	2
	Нездоровый для чувств.	10	56	16
	Опасный	2	42	75

В качестве визуализации процесса обучения на рисунке 8 представлены графики изменения точности (Accurancy) и функции потерь (Loss) в зависимости от номера эпохи обучения (epoch) сети с параметрами S=4, P=4 для обучающей (train) и тестовой (validation) выборок (пример классификации на 2 класса с сезонами).

Из графика видно, что процесс обучения плавно приближается к ожидаемому выходу, для достижения условия “функция потерь не изменяется более чем на 0,01 на протяжении 20 эпох” [6] потребовалось 35 эпох. Минимальная функция потерь на обучающих данных – 0,025. Максимальная точность на обучающих данных примерно

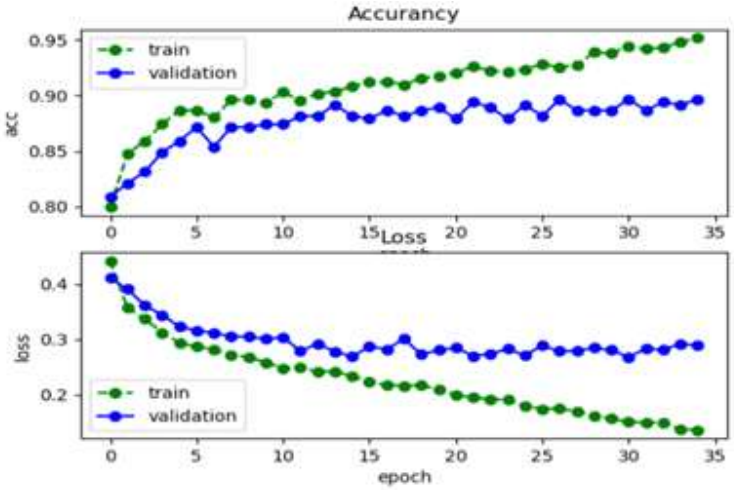


Рис. 8. Процесс обучения LSTM-сети с введенным режимом сезонности с параметрами S=4 и P=4

95%. Сравнивая эти показатели с полученными ранее [4], следует признать, что изменения в обработке данных показали свою эффективность.

Классификация AQI на 4 интегрированных класса. Дополнение наблюдений AQI за период с 23 ноября 2020 года по 31 марта 2021 (таблица 4, рисунок 9(а)) позволило рассмотреть задачу прогноза AQI как задачу классификации на 4 класса. Из анализа

количества наблюдений следует, что по-прежнему недостаточно наблюдений для прогноза классов “Очень нездоровый” и “Опасный”. Поэтому эти классы интегрированы с классом «Нездоровый» в один класс «Опасный» (рисунок 9(б)). При таком распределении данных, если всегда предсказывать “Умеренный” класс, то точность такого «наивного» прогноза составит примерно 49,4%. Значит, классификатор, прогнозирующий с точностью >50%, можно считать эффективным. У этого классификатора 4 выхода: вероятности классов «Хороший», “Умеренный”, “Нездоровый для чувствительных групп” и “Опасный”.

Таблица 4. Распределение AQI по классам для объединенного периода наблюдений 06/02/19–31/03/20 + 01/08/20–31/03/21.

Класс AQI	Количество наблюдений	Относительная доля
Хороший	865	0,151
Умеренный	2825	0,494
Нездоровый для чувствительных групп	762	0,133
Нездоровый	844	0,148
Очень нездоровый	275	0,048
Опасный	151	0,026

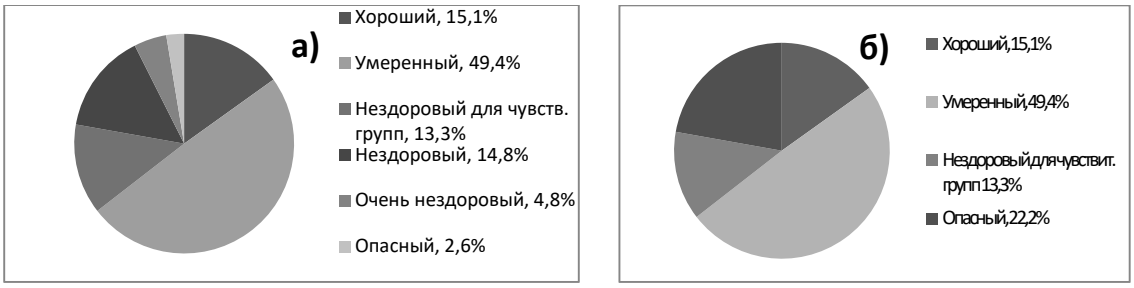


Рис. 9. Распределение наблюдений по классам AQI (а) и распределение наблюдений по трем интегрированным классам AQI (б) для объединенного периода 06/02/2019– 31/03/2020 и 01/08/2020 – 31/03/2021

В таблице 5 представлены результаты классификации AQI на 4 класса. Классификатор демонстрирует вполне приемлемую точность прогноза. Наилучшая точность F=0,8 соответствует глубине прогноза P=8 (то есть на одни сутки вперед), при этом необходимая «история» наблюдений S=16 (двое суток назад). Также в таблице приведены результаты классификации на 3 класса, соответствующих описанным в предыдущем параграфе, но на новой выборке данных. Как видно, наилучшую точность F=0,777 при этом показал вариант прогноза с глубиной P=4 (12 часов вперед) при истории наблюдений S=12 (36 часов назад). Повышение точности прогноза по сравнению с результатами, представленными в таблице

2, вполне объяснимо увеличением выборки наблюдений. Оба классификатора (и на 3 и на 4 класса) позволяют выполнять прогноз AQI на 4 суток вперед с точностью F≅0.77.

Таблица 5. F-мера точности модели прогнозирования четырех и трех классов AQI в зависимости от длины последовательности входных векторов (*S*) и глубины прогноза (*P*) для выборки наблюдений 06/02/2019–31/03/2020+01/08/2020–31/03/2021.

S	**P**	**F-мера 3 класса 06/02/2019-31/03/2020+ 01/08/2020-31/03/2021**	**F-мера 4 класса 06/02/2019-31/03/2020+ 01/08/2020-31/03/2021**
8	4	0.7596	0.7641
8	8	0.7359	0.7449
8	16	0.7110	0.7286
8	24	0.7341	0.7270
8	32	0.7317	0.7139
12	4	**0.7771**	0.7669
12	8	0.7591	0.7728
12	16	0.7460	0.7571
12	24	0.7628	0.7556
12	32	0.7591	0.7456
16	4	0.7733	0.7895
16	8	0.7685	**0.8007**
16	16	0.7601	0.7773
16	24	0.7530	0.7772
16	32	0.7684	0.7733

Классификатор на 4 класса демонстрирует также приемлемые обучающие свойства. На рисунке 10 представлены графики изменения точности (Accurancy) и функции потерь (Loss) сети с параметрами S=16, P= 8. Заметим, что процесс обучения вполне отвечает рекомендациям, приведенным в [6].

Заключение. Таким образом, в работе показаны результаты разработки классификатора на 3 класса AQI (“Умеренный”/“Нездоровый для чувствительных групп” /“Опасный” (AQI≤100 / 100<AQI≤150/ AQI>150)) и классификатора на 4 класса AQI (“Хороший”/“Умеренный”/“Нездоровый для чувствительных групп” /“Опасный” (AQI≤50/ 50<AQI≤100 / 100<AQI≤150/ AQI>150)). Используемая при разработке выборка данных объединяет два периода наблюдений: 06/02/19–31/03/20 и 01/08/20–31/03/21. Прогноз AQI на 3 и 4 класса

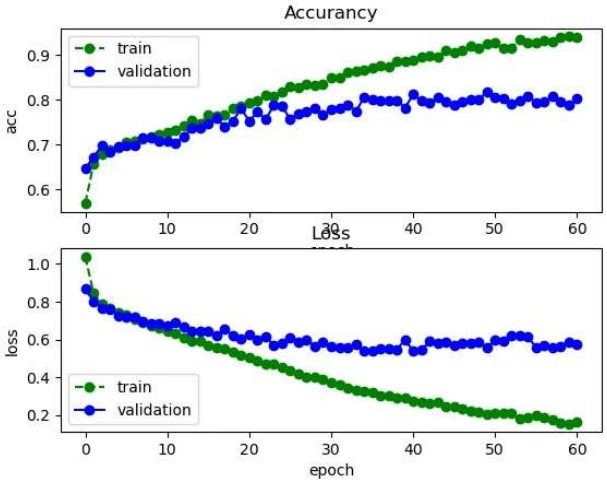


Рис.10. Процесс обучения LSTM-сети с параметрами S=16 и P=8, классифицирующей AQI на 4 класса

возможен до 4 дней вперед с точностью ≅77%. Дальнейшее увеличение истории наблюдений позволит прогнозировать классы с AQI >150.

Литература

1. AirNow Depatment of State // https://airnow.gov/index.cfm?action=airnow.global_summary #U.S._Department_of_State $Bishkek, (дата обращения: 30.04.2021).
2. Лыченко Н.М., Великанова Л.И., Верзунов С.Н., Сороковая А.В. Модели прогноза уровня загрязнения атмосферного воздуха г. Бишкека// Вестник КРСУ. – Т.24. – №4. – 2021. –С.87–95.
3. Air Quality Index (AQI) - A Guide to Air Quality and Your Health. US EPA. 9 December 2011.
4. Лыченко Н.М., Сороковая А.В. Применение LSTM-нейронных сетей для классификации индекса качества воздуха г. Бишкека // Проблемы автоматики и управления. 2020. – № 1 (38). – С. 70–80. DOI: 10.5281/zenodo.3904130
5. Understanding LSTM Networks. – URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. (дата обращения 14.09.2020)
6. Курс CS231n: Convolutional Neural Networks for Visual Recognition. - URL: https://cs231n.github.io/neural-networks-3/#baby (дата обращения 01.09.2020)
7. Сайт «Расписание погоды rp5.ru» Архив погоды в Бишкеке https://rp5.ru/%D0%90%D1%80%D1%85%D0%B8%D0%B2_%D0%BF%D0%BE%D0%B3%D0%BE%D0%B4%D1%8B_%D0%B2_%D0%91%D0%B8%D1%88%D0%BA%D0%B5%D0%BA%D0%B5 (дата обращения: 30.04.2021)
8. Оценка классификатора (точность, полнота, F-мера). - URL: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html. (дата обращения 29.05.2020).