УДК 004.8, 004.93,159.95

ФОРМИРОВАНИЕ ПРОСТРАНСТВА ПРИЗНАКОВ ДЛЯ ОБНАРУЖЕНИЯ ЖИВЫХ ОБЪЕКТОВ В ЗДАНИИ НА ОСНОВЕ ЭКОЛОГИЧЕСКИХ ФАКТОРОВ

© 2016 И.М. Куликовских

Самарский национальный исследовательский университет имени академика С.П. Королёва

Статья поступила в редакцию 11.11.2016

В работе рассматривается задача формирования признакового описания для обнаружения живых объектов на основе экологических факторов. Для решения поставленной задачи была реализована модель логистической регрессии и предложен функционал, учитывающий взаимную корреляцию признаков. Серия вычислительных экспериментов подтвердила адекватность и непротиворечивость полученных результатов, а также эффективность предложенной модели для обнаружения объектов в здании.
Ключевые слова: машинное обучение, формирование пространства признаков, логистическая регрессия, бинарная классификация, обнаружение объектов, экологические факторы

Работа выполнена при государственной поддержке Министерства образования и науки РФ (грант № 074-U01).

ПОСТАНОВКА ЗАДАЧИ

Проблема обнаружения живых объектов в здании является актуальной для энергосбережения и обеспечения безопасности в помещениях. Как показывают результаты предыдущих исследований [1-7], более точное решение данной проблемы связано с анализом экологических факторов, что позволило повысить энергосбережение с 30% до 42% [1-3]. С другой стороны такие системы точного обнаружения позволяют определить поведение и перемещение живых объектов без использования камеры, что представляет значительный интерес из-за необходимости соблюдения конфиденциальности информации.

Одним из способов повышения точности обнаружения является применение методов машинного обучения для анализа поступающей с датчиков информации, что получило широкое применение [3-7]. В работе [5] анализировался уровень CO_2внутри и снаружи помещения, который затем использовался для построения марковских моделей, нейронных сетей и метода опорных векторов. В другом исследовании [6] строились деревья решений на основе информации с датчиков CO_2, света, звука и пассивных инфракрасных датчиков. В работе [7] формировалась информация по свету, звуку, состоянию датчика с язычковым контактом, CO_2, температуре, состоянию пассивных инфракрасных датчиков для построения нейронных сетей с радиальной базисной функцией. Наконец, в [4] анализировались уровни CO_2, влажности, температуры и света с применением случайного леса, CARTмодели и линейного дискриминантного анализа (LDA). Представленные в последней работе результаты показали наибольшую точность для метода LDA по сравнению с результатами аналогичных исследований.

В свою очередь, логистическая регрессия имеет ряд достоинств по сравнению с LDA [8-14], в частности, даёт лучшие результаты, поскольку основана на менее жёстких гипотезах. Кроме того, логистическая регрессия предпочтительнее, так как не вводит избыточную сущность как LDA, который сводит задачу классификации к более сложной задаче восстановления плотностей вероятностей [8].

Однако в работе[4]отмечается, что реализация модели логистической регрессии на используемом наборе данных невозможно, так как алгоритм расходится для линейно разделимых классов. Следовательно, было бы интересно оценить возможности логистической регрессии в контексте задачи, описанной в [4], более детально, проанализироватьуказанное ограничение и попытаться реализовать модель логистической регрессии при решении задачи обнаружения живых объектов в здании.

ОПИСАНИЕ НАБОРА ДАННЫХ

Исходные наборы данных для решения поставленной задачи доступны в UCI Machine Learning Repository по ссылке https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+. Данные наборы могут быть использованы как для обучения, так и для тестирования моделей. Для сбора данных было использовано помещение 5,85 × 3,50 × 3,53 м, в котором были измерены следующие экологические факторы: уровни температуры, влажности, света и CO_2, с помощью датчиков, установленных в помещении. Кроме того, для формирования меток – отсутствия или присутствия живого объекта в

Куликовских Илона Марковна, кандидат технических наук, доцент кафедры информационных систем и технологий. E-mail: kulikovskikh.i@gmail.com

помещении – была установлена цифровая камера, которая снимала изображения с заданным интервалом времени. Таким образом, представленные выше признаки, дополнены отсчетами времени и влагоемкостью. Результирующий показатель в наборах соответствует статусу – обнаружен, не обнаружен – и определяет метки классов. Более подробное описание процедуры сбора данных представлено в работе [4].

Обучающая выборка содержит 8143 реализации, которые были получены, когда дверь в помещение была закрыта. Два тестовых набора по 1998 реализаций каждая были сформированы для двух случаев: открытой и закрытой двери. На рисунке ниже представлены исходные наборы в зависимости от имеющихся признаков, исключая отсчеты времени. Метки классов размечены в виде черных крестиков в случае отсутствия объекта и красных точек в случае его присутствия в помещении (см. рис. 1).

Исходя из описания данных, представим задачу обнаружения объектов в здании как задачу бинарной классификации исходного набора данных с формированием пространства признаков. В качестве метода классификации определим нереализованный в проведенных ранее исследованиях [4] метод логистической регрессии и проанализируем его эффективность.

ФОРМИРОВАНИЕ ПРОСТРАНСТВА ПРИЗНАКОВ

Приведем основные понятия и определения, необходимые для решения поставленной задачи обнаружения объектов с помощью модели логистической регрессии.

Определение 1. *Пусть X – множество объектов, Y – множество допустимых ответов. Объекты описываются числовыми признаками* $f_j: X \to \mathbb{R}$, $j = \{1, n\}$, *где n – количество признаков. Тогда в русле работы* [8] *вектор* $(x^j)_{j=1}^n \in \mathbb{R}^n$, *где* $x^j = f(x)$, *называется пространством признаков объекта x.*

Определение 2. *Пусть X – множество объектов,* Y *– множество допустимых ответов, а* Θ *– множество допустимых значений пространства параметров* θ. *Тогда в русле работы* [10] *параметрическим семейство* $A = \{g(x, \theta) | \theta \in \Theta\}$, *где* $g: X \times Y \to Y$ *– фиксированная функция, называется моделью алгоритмов.*

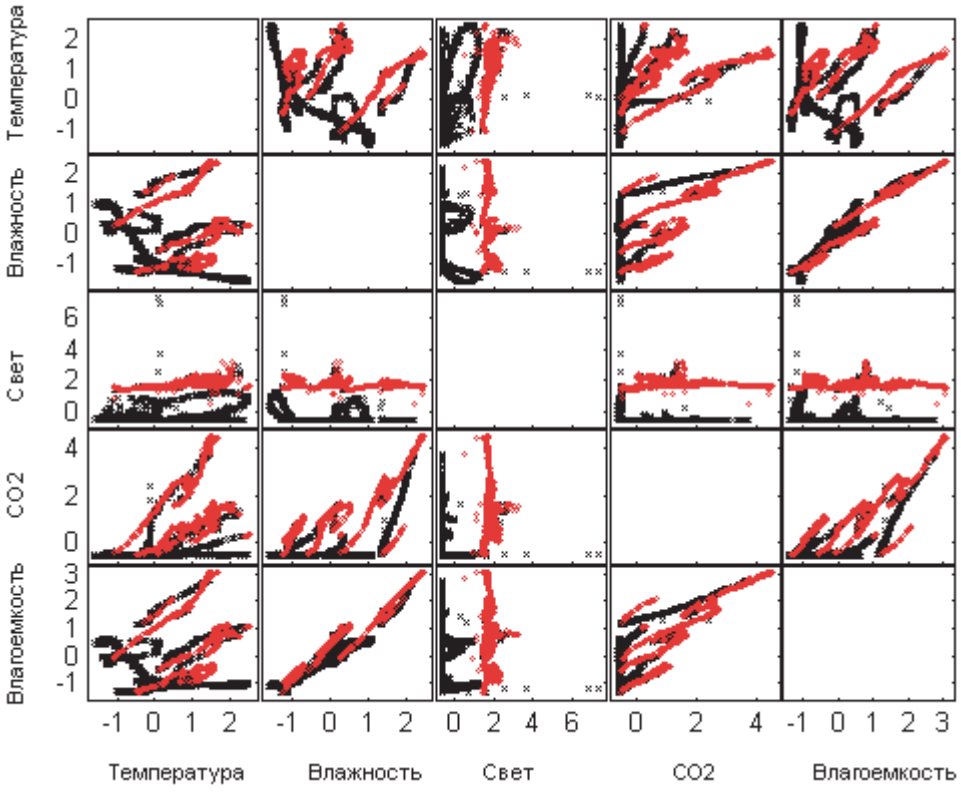
Задача 1. *Пусть в качестве модели алгоритмов* $a \in A$ *выбрана модель логистической регрессии*

$$g(x, \theta) = \frac{1}{1 + \exp(-\theta^T x)}.$$

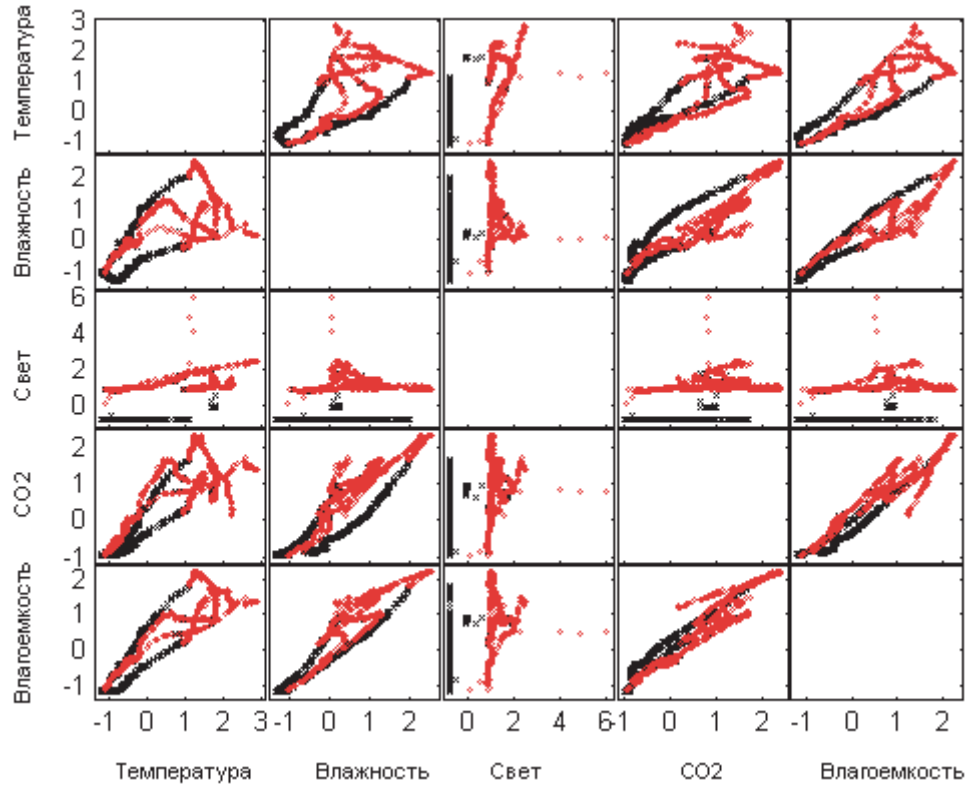
Тогда задача определения пространства параметров $\theta \in \Theta$ *по выборке прецедентов* $X^l = (x_i, y_i)_{i=1}^l$, *где* $y_i \in \{0,1\}$, *сводится к минимизации логарифмической функции потерь* $\ln L(\theta, X^l)$ [8]

$$\ln L(\theta, X^l) = \sum_{i=1}^{l} \ln\left(1 + \exp(-\theta^T x_i y_i)\right).$$

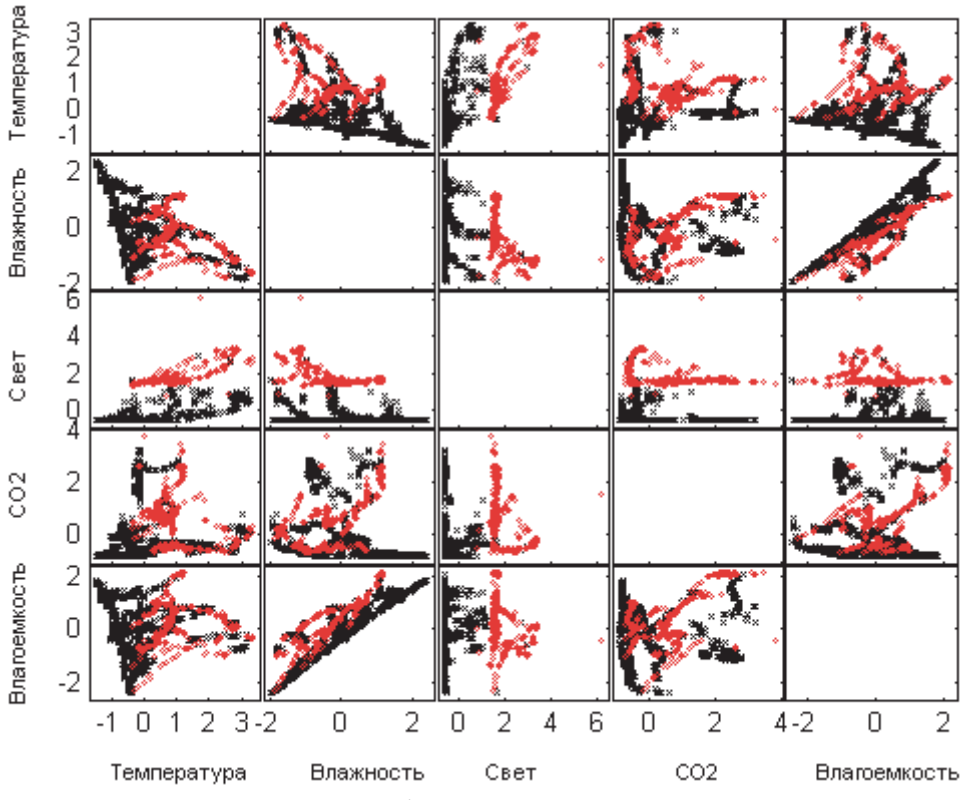
При этом в постановке *Задачи 1* значение y_i=*0* соответствует случаю отсутствия объекта, а y_i=*1* – случаю присутствия объекта.



а)



б)



в)

Рис. 1. Исходные данные:
а – обучающая выборка; б – тестовая выборка (закрытая дверь); в – тестовая выборка (открытая дверь)

Сформируем пространство признаков с учетом взаимной корреляции признаков. С этой целью построим таблицу корреляций признаков для обучающей выборки $X^{l}=(x_i,y_i)^{l}_{i=1}$ и тестовых выборок $X^{k_1}=(x_i,y_i)^{k_1}_{i=1}$ и $X^{k_2}=(x_i,y_i)^{k_2}_{i=1}$. Согласно структуре исходных данных l=6107, k_1=1998, k_2=1998; n=5, x^j = {Температура, Влажность, Свет, CO_2, Влагоёмкость}. Заметим, что часть данных из обучающей выборки было использовано для проведения кросс-валидации.

Зададим функционал для формирования пространства признаков $x^j = f(x)$ для пары $\{x^p, x^q\}$, где

Таблица 1. Коэффициенты корреляции между признаками x^j для выборок X^l, X^{k_1} и X^{k_2}

	$x^j \subset X^l$					$x^j \subset X^{k_1}$					$x^j \subset X^{k_2}$				
j	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	1,00	-	-	-	-	1,00	-	-	-	-	1,00	-	-	-	-
2	-0,15	1,00	-	-	-	0,71	1,00	-	-	-	-0,48	1,00	-	-	-
3	**0,65**	0,04	1,00	-	-	0,77	0,56	1,00	-	-	0,70	-0,2	1,00	-	-
4	**0,56**	0,44	**0,66**	1,00	-	0,87	0,91	0,77	1,00	-	0,22	-0,1	0,23	1,00	-
5	0,15	**0,96**	0,23	**0,63**	1,00	0,89	0,95	0,70	0,96	1,00	-0,03	0,88	0,15	0,05	1,00

$\{p, q\} \subset j,$, в следующем виде

$$f(x^p, x^q, d) = \prod_{m=1}^{d} \prod_{n=0}^{m} (x^p)^{m-n} (x^q)^n,$$

где $d = 2^k, k = \overline{0,3}$. При этом в данном исследовании выбиралась пара $\{p,q\}$, соответствующая признакам с наибольшей корреляцией на обучающей выборке $x^j \subset X^l$ (см. табл. 1).Исключением являлась пара {*2,5*}, которая линейно зависима по определению [4].

Рассмотрим различные комбинации признаков $f(x^p, x^q, d)$ в виде наборов:

1. G_6: $f(x^1, x^3, d) + f(x^1, x^4, d) + f(x^3, x^4, d) + f(x^4, x^5, d)$;
2. G_5: $f(x^3, x^4, d) + f(x^4, x^5, d)$;
3. G_4: $f(x^1, x^3, d) + f(x^3, x^4, d)$
4. G_3: $f(x^4, x^5, d)$;
5. G_2: $f(x^3, x^4, d)$;
6. G_1: $f(x^1, x^3, d)$

РЕЗУЛЬТАТЫ РАСПОЗНАВАНИЯ

Проанализируем приведенные выше комбинации и исследуем возможность минимизации набора признаков. В таблице 2 приведены значения точности распознавания для наборов $\{Gr\}_{r=1}^{6}$ при различных значениях величины d = {2,4,6} на обучающей выборке и двух тестовых выборках. Следует отметить, что при формировании наборов была проведена нормировка значений для повышения качества и скорости сходимости выбранной модели классификации [8]. Нормировка пространства признаков также необходима из-за выбора степенного функционала $f(x^p, x^q, d)$ – возведение в степень слишком больших или слишком маленьких значений может привести к неадекватным результатам. Модель логистической регрессии была реализована в системе GNU Octave 3.8.2 и апробирована при проведении экспериментальных исследований на MacBookAir 11 OS X EI Captain с процессором 1.3 GHzIntelCore i5 и памятью 4 GB 1600 MHz DDR3.

Как видно из таблицы, повышение сложности пространства признаков не приводит к существенному повышению качества распознавания: точность классификации как среднее лучших показателей (выделено жирным) при d = 2 на 0,85% хуже, чем при d = 4, и на 0,25% хуже, чем при d = 6. Более того, обращает на себя внимание тот факт, что при d = *{2,4}* наилучшая точность получена для результатов, где при формировании признаков использовался функционал $f(x^1,x^3,d)$ с парой признаков *{1,3}*.

Согласно результатам исследований, приведенных в [4], хорошее качество классификации было получены приполном наборе G_0:x^j – 97,90% для тестового набора X^{k_1} и 98,76% для тестового набора X^{k_2}.Данные результаты были получены с помощью LDA.Тем не менее, отмечается, что наилучший результат для набора X^{k_1} – **97,9%** был также получен на паре признаков {1,3}, а для набора X^{k_2} – **99,33%** на полном наборе, но дополненном парой вновь сформированных признаков, учитывающих временную компоненту. Представленные результаты были тоже получены с использованием LDA.

Анализируя таблицу 1, можно заметить, что пара признаков {1,3} является единственной, которая имеет высокую корреляцию как всех выборках: X^l, X^{k_1} и X^{k_2}. Следовательно, использование корреляции при формировании пространства признаков является целесообразным. Проведем серию дополнительных вычислительных экспериментов при d = 1, включив в рассмотрение полный набор признаков x^j без дополнительного преобразования. Кроме того, для сравнения

Таблица 2. Точность распознавания при формировании наборов $\{Gr\}_{r=1}^{6}$ при d = {2,4,6}

	$d = 2$			$d = 4$			$d = 6$		
	X^l	X^{k_1}	X^{k_2}	X^l	X^{k_1}	X^{k_2}	X^l	X^{k_1}	X^{k_2}
G_6	98,4608	92,8929	85,8359	99,1158	**94,5445**	89,0891	98,9520	90,8408	90,9910
G_5	98,4117	91,1411	89,0891	98,8701	94,4444	82,8829	98,6900	91,8919	**97,9980**
G_4	98,7392	**93,2933**	93,8438	99,1158	93,1932	91,0410	99,1813	89,8899	92,5926
G_3	93,6794	73,7738	69,8699	95,2186	74,0240	54,8549	94,9894	75,7758	65,5656
G_2	98,5754	91,7918	96,1962	98,8374	94,4444	92,2422	98,6573	**92,9429**	94,9450
G_1	98,6573	75,5756	**97,1471**	99,0175	76,4764	**97,5976**	98,9029	76,3763	94,8949

Таблица 3. Точность распознавания при формировании наборов $\{G_r\}^6_{r=0}$ при $d = 1$

	С нормировкой			Без нормировки		
	X^l	X^{k_1}	X^{k_2}	X^l	X^{k_1}	X^{k_2}
G_6	98,4444	87,9379	94,3443	98,4444	97,3974	96,9469
G_5	98,6737	89,5896	97,0971	98,6900	97,8478	99,1992
G_4	98,6737	88,738739	97,1972	98,6737	97,9980	99,2492
G_3	92,6314	79,2793	70,7207	91,0431	86,9369	77,2773
G_2	98,6900	89,7397	97,1471	98,7064	97,8979	99,1992
G_1	98,8701	79,3293	**99,3493**	98,8538	**97,9980**	**99,3494**
G_0	94,0888	**96,9970**	85,8358	98,6737	97,8979	99,1992

результатов с результатами, представленными в [4], проанализируем качество классификации как с нормировкой, так и без нормировки пространства признаков (см. табл. 3).

Из представленной таблицы следует, что точность распознавания без нормировки пространства признаков значительно выше. Результаты, полученные в данном исследовании с помощью логистической регрессии – **98%** для набора X^{k_1} и 99,35% для набора X^{k_2},– аналогичны представленным в [4] для LDA, но реализуют более простую и легко интерпретируемую модель.

ВЫВОДЫ

В данной работе:

1. реализована модель логистической регрессии, являющаяся более простой и интерпретируемой по сравнению с рассмотренными ранее;

2. предложен метод обучения классификатора на основе формирования признаков с учетом взаимной корреляции и выявлена пара наиболее информативных признаков.

БЛАГОДАРНОСТИ

Автор выражает благодарность д.т.н., профессору С.А. Прохорову и к.ф.-м.н., профессору Л.П. Усольцеву за ценные замечания и рекомендации, способствующие повышению качества представления результатов исследований.

СПИСОК ЛИТЕРАТУРЫ

1. *Erickson V.L., Carreira-Perpiñán M.Á., Cerpa A.E.* OBSERVE: Occupancy-based system for efficientreduction of HVAC energy // Information Processing in Sensor Networks (IPSN): Proc. 10thIEEE International Conference on, Stockholm, Sweden, 2011. Pp. 258-269.
2. Occupancy modeling and prediction for building energy management / *V.L. Erickson, M.Á. Carreira-Perpiñán, A.E. Cerpa* // ACM Transactions on Sensor Networks (TOSN). 2014. 10(3). 42.
3. *Dong B., Andrews B.* Sensor-based occupancy behavioral pattern recognition for energy and comfortmanagement in intelligent buildings. URL: www.ibpsa.org/proceedings/BS2009/BS09_1444_1451.pdf (дата обращения 8.11.2016).
4. *Candanedo L.M., Feldheim V.* Accurate occupancydetection of an office room from light, temperature, humidity and CO2measurements using statistical learning models // Energy and Buildings. 2015. URL: http://dx.doi.org/10.1016/j.enbuild.2015.11.071 (дата обращения 8.11.2016).
5. Occupancy detection through an extensive environmental sensor network in an open-plan office building / *K.P. Lam, M. Höynck, B. Dong, B. Andrews, Y.-S. Chiou, R. Zhang, D. Benitez, J. Choi* // IBPSA Building Simulation. 2009. 145. pp. 1452-1459.
6. Real-time occupancy detection using decision trees with multiple sensor types / *E. Hailemariam, R. Goldstein, R. Attar, A. Khan* // Simulation for Architecture and Urban Design: Proc. 2011 Symposium on, Boston, MA, USA, 2011. pp. 141-148.
7. A multi-sensor based occupancy estimation model forsupporting demand driven HVAC operations // *Z. Yang, N. Li, B. Becerik-Gerber, M. Orosz* // Simulation for Architecture and Urban Design: Proc. 2012 Symposium on, San Diego, CA, USA, 2012. pp. 49-56.
8. *Воронцов К.В.* Лекции по линейным алгоритмам классификации. URL: http://www.machinelearning.ru/wiki/images/6/68/voron-ML-Lin.pdf Дата обращения 08.11.16.
9. *Воронцов К.В.* Лекции по алгоритмам восстановления регрессии. URL: http://www.ccas.ru/voron/download/Regression.pdf. Дата обращения 08.11.16.
10. *Воронцов К.В.* Математические методы обучения по прецедентам (теория обучения машин). URL: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf Дата обращения08.11.16.
11. *Rodriguez G.* Lecture notes on generalized linear models. Appendix B. Generalized linear model theory. URL: http://data.princeton.edu/wws509/notes/a2.pdf. Accessed 08.11.2016.
12. *Rodriguez G.* Lecture notes on generalized linear models. Chapter 3. Logit models for binary data. URL: http://data.princeton.edu/wws509/notes/c3.pdf. Accessed 08.11.2016.
13. *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning: Data mining, inference, and

prediction (2nd ed.)/T. Hastie, Springer Series in Statistics, 2013. 745 p.
14. *Czepiel S.A.* Maximum likelihood estimation of logistic regression models: Theory and implementation. URL: http://czep.net/stat/mlelr.pdf. Accessed 08.11.2016.

FEATURE EXTRACTION TO DETECT OCCUPANCY IN BUILDINGS USING ECOLOGICAL FACTORS

Samara National Research University named after Academician S.P. Korolyov

The paper delves into feature extraction problem to detect occupancy in buildings using ecological factors. To solve this problem a logistic regression model was implemented and a composed function was proposed. This composed function took into account features cross-correlations. The computational experiments confirmed the adequacy and consistency of research results as well as the efficiency of created models to detect occupancy in buildings.
Keywords: machine learning, feature extraction, logistic regression, binary classification, occupancy detection, ecological factors

Ilona Kulikovskikh, PhD, Associate Professor of Information System and Technologies Department.
E-mail: kulikovskikh.i@gmail.com