Мир науки. Социология, филология, культурология https://sfk-mn.ru
World of Science. Series: Sociology, Philology, Cultural Studies

2021, №4, Том 12 / 2021, No 4, Vol 12 https://sfk-mn.ru/issue-4-2021.html

URL статьи: https://sfk-mn.ru/PDF/21SCSK421.pdf

Ссылка для цитирования этой статьи:
Самойлова, Т. А. Сравнительное исследование эффективности алгоритмов классификации данных в социологическом анализе цветонаименований / Т. А. Самойлова, Ю. А. Грибер // Мир науки. Социология, филология, культурология. — 2021. — Т. 12. — № 4. — URL: https://sfk-mn.ru/PDF/21SCSK421.pdf

For citation:
Samoylova T.A., Griber Yu.A. A comparative study of the effectiveness of data classification algorithms in sociological analysis of color names. *World of Science. Series: Sociology, Philology, Cultural Studies*, 4(12): 21SCSK421. Available at: https://sfk-mn.ru/PDF/21SCSK421.pdf. (In Russ., abstract in Eng.).

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 17-29-09145 «Картография цвета: диагностика развития цветонаименований русского языка с использованием естественно-научных, историографических, социологических и психологических методов»

УДК 316.77

ГРНТИ 04.51.51

Самойлова Татьяна Аркадьевна
ФГБОУ ВО «Смоленский государственный университет», Смоленск, Россия
Доцент кафедры «Информатики»
Кандидат технических наук, доцент
E-mail: tatsamoilova24@gmail.com
РИНЦ: https://elibrary.ru/author_profile.asp?id=100995

Грибер Юлия Александровна[1]
ФГБОУ ВО «Смоленский государственный университет», Смоленск, Россия
Профессор кафедры «Социологии и философии», директор «Лаборатории цвета»
Доктор культурологии
E-mail: y.griber@gmail.com
ORCID: https://orcid.org/0000-0002-2603-5928
РИНЦ: https://elibrary.ru/author_profile.asp?id=303167
Researcher ID: https://www.researcherid.com/rid/AAG-4410-2019
SCOPUS: https://www.scopus.com/authid/detail.url?authorId=56809444600

Сравнительное исследование эффективности алгоритмов классификации данных в социологическом анализе цветонаименований

Аннотация. В статье представлено сравнительное исследование различных алгоритмов классификации для прогнозирования пола респондента по его ответам в онлайн-эксперименте, направленном на изучение социальной дифференциации системы цветонаименований русского языка. Материалом исследования стали данные онлайн-эксперимента (http://colournaming.com), в котором в 2018–2020 годах приняли участие 2457 носителей русского языка (1402 женщины, 1055 мужчин), принадлежащих к разным возрастным группам в диапазоне от 16 до 98 лет (средний возраст — 41.36 лет, SD = 17.71). Каждый из полученных в ходе исследования ответов (N = 55515) содержал ряд принципиально различных по характеру признаков, фиксировавших не только координаты цветового образца в системе CIELAB и присвоенное ему

[1] Сайт лаборатории: http://color-lab.org/

цветонаименование (простое или сложное слово, словосочетание), но и социально-демографическую информацию о поле и возрасте респондента, месте его рождения и постоянного проживания, уровне образования и профессии. Авторы анализируют различные алгоритмы классификации с использованием программных библиотек NumPy, Pandas, Scikit-learn для языка программирования Python. Эффективность классификаторов оценивается по таким параметрам, как точность, полнота, F-мера и кривая ошибок. Результаты моделирования показывают, что алгоритм дерева решений классифицирует данные с точностью 92 % и качеством, соответствующим значению AUC = 0,99. Это значит, что именно его лучше всего использовать в обработке и анализе полученных данных. Представленная в работе методика оценки эффективности алгоритма классификации с использованием комплекса взаимодополняющих метрик может использоваться в качестве модели для выбора наиболее подходящего программного средства с учетом специфики конкретного случая в дальнейших социологических исследованиях.

Ключевые слова: эксперимент; социологический анализ данных; социальная дифференциация языка; цветонаименования; машинное обучение; классификация; Phyton

Введение

Принципиальными особенностями собираемых данных в исследованиях коллективных реакций на цвет представителей различных социальных групп является их большой объем, с одной стороны, и необычный характер — с другой. Если преимущественно социология имеет дело с числовыми данными, которые хорошо поддаются анализу традиционными методами математической статистики и вычислительной математики, то в социологии цвета данные имеют принципиально другой характер: они могут представлять собой разные по длине тексты (см. напр.: [1; 2]), изображения (см. напр.: [3–5]) и даже видеофайлы (см. напр.: [6; 7]).

В условиях, когда традиционные методы анализа данных не способны обеспечить получение результата с необходимой точностью за приемлемое время, особенно важным становится вопрос выбора и применения программных средств с учетом специфики конкретного случая социологического исследования (см. подр. об этом: [8; 9, с. 56–58; 10, с. 355–367]).

Цель работы заключается в том, чтобы провести сравнительное исследование различных алгоритмов классификации и определить лучший классификатор для прогнозирования пола респондента по его ответам в онлайн-эксперименте, направленном на изучение социальной дифференциации системы цветонаименований русского языка.

Материал исследования

Материалом исследования стали данные онлайн-эксперимента (http://colournaming.com), в котором в 2018–2020 годах приняли участие 2457 носителей русского языка (1402 женщины, 1055 мужчин), принадлежащих к разным возрастным группам в диапазоне от 16 до 98 лет (средний возраст — 41.36 лет, SD = 17.71). Задача участников заключалась в том, чтобы подобрать наиболее подходящее название для набора виртуальных цветовых стимулов, которые компьютер в случайном порядке отбирал из 606 экспериментальных образцов (см. подр.: [11–13]). Каждый из полученных в ходе исследования ответов (N = 55515) содержал ряд различных по характеру признаков, фиксировавших не только координаты цветового образца в системе CIELAB и присвоенное ему цветонаименование (простое или сложное слово, словосочетание), но и социально-

демографическую информацию о поле и возрасте респондента, месте его рождения и постоянного проживания, уровне образования и профессии.

Метод исследования

Гипотеза исследования состояла в том, что система цветонаименований русского языка социально дифференцирована и имеет разновидности, обусловленные социальным расслоением носителей. Вариативность системы цветонаименований, которая обнаруживается в языковых различиях между представителями социальных слоев и социальных групп, не просто предполагает наличие в языке альтернативных социально обусловленных цветонаименований, она программирует переописание структуры, плотности, состава и свойств этой системы, в результате чего формируются различные денотативные карты цветового поля, происходит смещение центроидов фокальных цветов и изменяется частотность использования цветообозначений.

Поскольку, согласно имеющимся данным, среди всех социальных характеристик носителей наиболее сильные корреляции различной структуры цветонаименований отмечаются с полом (см. напр.: [13; 14]), мы решили проверить, сможет ли машина научиться отличать ответы мужчин от ответов женщин, действительно ли между ними и в наших данных обнаружится настолько большая разница.

Для решения этой задачи был выбран метод классификации (см., напр.: [15, с. 62–117]). Классификация представляет собой один из наиболее широко используемых типов задач принятия решений в алгоритмах машинного обучения. Основная цель классификации — точно предсказать целевой класс для определенного экземпляра данных. На этапе обучения модели классификатора каждый экземпляр данных имеет предопределенный целевой класс, тогда как на этапе тестирования неизвестные тестовые экземпляры прогнозируются с использованием обученной модели. Алгоритм классификации обрабатывает большой объем данных обучающей выборки и строит модель. Обработка тестовых данных предшествует классификации, для того чтобы улучшить качество данных.

Результаты и их обсуждение

Предварительная обработка исходных данных

В нашем случае предварительная обработка исходных данных включала их очистку и отбор наиболее полезных признаков. Очистка заключалась в удалении результатов экспермента с недопустимыми значениями атрибутов. Отбор признаков использовался для уменьшения размерности данных, предназначенных для дальнейшей классификации. При этом из набора данных были удалены ненужные и избыточные атрибуты, которые имели меньшее значение для классификации. Задачи предварительной обработки и очистки первоначальных данных выполнялись средствами VBA в среде Excel, в которой они хранились.

После преобразования окончательная база данных составила 48021 записей, включающих 6 числовых переменных — 5 входных и одну целевую. Входные переменные представляли собой хроматические характеристики цвета (показатели L*, a*, b* системы CIELAB), возраст участников (age) и присвоенное цвету имя (colorname). Переменная colorname включала три класса, соответствующих типу выбранного респондентом цветонаименования (1 — основной термин цвета: *красный, оранжевый, жёлтый, зелёный, голубой, синий, фиолетовый, розовый, коричневый, белый, серый, чёрный*; 2 — цветонаименование, производное от основного (например, *тёмно-красный*), 3 — все

остальные цветонаименования). Целевая переменная состояла из двух классов, соответствующих полу (0 — мужчина, 1 — женщина) (табл. 1).

Таблица 1

Характеристики переменных записи данных

Атрибут	Значения
L*	0–100
a*	-100–+100
b*	-100–+100
age	16–100
colorname	1–3
класс — gender	(0 — мужчина, 1 — женщина)

Составлена авторами на основе собственного исследования

Хорошие наборы данных содержат признаки, не коррелированные друг с другом, поскольку в машинном обучении большая степень корреляции приводит к снижению эффективности модели прогнозирования. Чтобы проверить отобранные признаки, мы рассчитали в Python корреляционную матрицу атрибутов набора данных с помощью функции corr () библиотеки Pandas. Такой анализ показал, что у всех пар отобранных атрибутов отсутствовала корреляция выше порогового значения 0.42 (рис. 1).

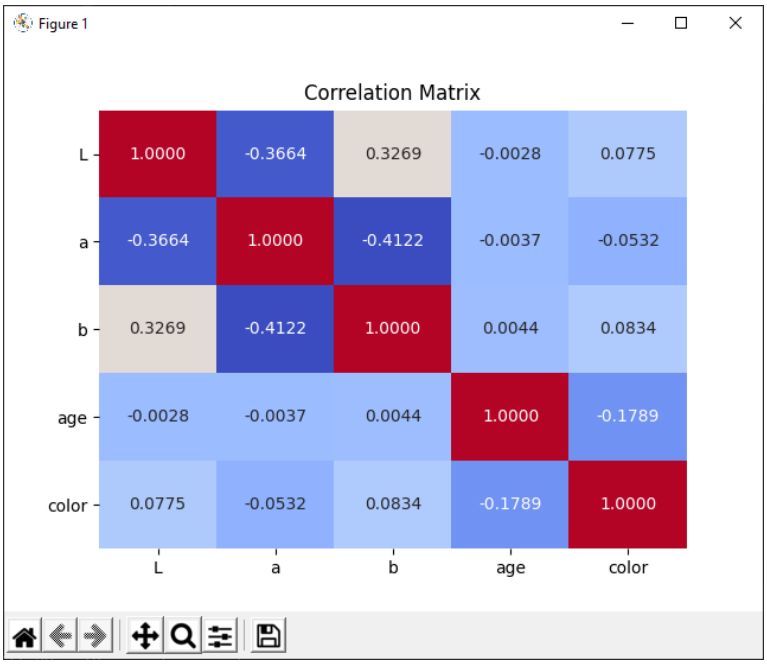

Рисунок 1. *Матрица корреляции отобранных признаков (рисунок авторов)*

Алгоритмы классификации

Чтобы выбрать наилучший алгоритм, мы применили пять популярных классификаторов для обучения набора социологических данных (см. подр.: [15–17]): (1) *Decision Tree* (строит дерево решений, используя энтропию и информационный прирост каждого атрибута; процесс заключается в последовательном, рекурсивном разбиении обучающего множества на подмножества с применением решающих правил в узлах); (2) *Gradient Boosting* (последовательность деревьев, каждое из которых улучшает прогноз предыдущего с использованием остаточного градиентного повышения); (3) *Random Forest* (случайным образом создает деревья решений на выбранных выборочных данных и получает прогноз для каждого дерева); (4) *KNeighbors* (работает по принципу, предполагающему, что каждая точка

данных, расположенная рядом с соседями, относится к тому же классу) и (5) *GaussianNB* (вариант наивного байесовского метода, который следует нормальному распределению по Гауссу).

Оценка точности и производительности классификатора

Все модели строились с использованием перекрестной проверки (cross-validation), для которой весь набор данных был случайным образом разделен на две части: обучающий состоял из 2/3 данных (32174) для разработки модели, а тестовый — из оставшейся 1/3 данных (15847) для ее проверки. Для выбора наиболее эффективного алгоритмы сравнивались на основе точности (Accuracy), прецизионности (Precision), отзыва (Recall) F-меры (f1-score, индекс комплексной оценки) и площади под кривой ошибок.

Точность, которая представляет собой процентную меру правильно классифицированных экземпляров для всех экземпляров, рассчитывалась по формуле (1):

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

Прецизионность — правильно классифицированные экземпляры для тех экземпляров, которые классифицируются как положительные, вычислялась по формуле (2):

$$Precision = TP/(TP + FN) \quad (2)$$

Эта метрика хорошо дополняет меру точности в задачах с неравными классами, к которым относится наша, где число испытуемых мужчин значительно меньше числа испытуемых женщин.

Отзыв — мера положительного экземпляра, который правильно классифицирован, рассчитывалась с помощью метрики (3):

$$Recall = TP/(TP + FP) \quad (3)$$

Для расчета F1-score — комбинированного показателя точности и полноты, гармонического среднего обоих — использовалась формула (4):

$$F1\text{-}score = 2*Precision*Recall/(Precision + Recall) \quad (4)$$

Рассматриваемые метрики эффективности алгоритмов основаны на использовании следующих исходов классификации: истинно положительные (TP), истинно отрицательные (TN), ложноположительные (FP) и ложноотрицательные (FN). Эти исходы рассчитываются по результату работы модели, заключающемся в определении пола — женщина (тогда результат = true) или мужчина (тогда результат = false). Если модель верно определила женщину и поставила 1 (true, положительный класс), тогда это истинно положительный исход (TP), если же модель ставит женщине 0 (false, отрицательную метку), тогда это ложно отрицательный исход (FP). Для мужчин аналогичными значениями являются TN и FN соответственно.

Результаты вычисления точности, прецизионности, отзыва, *F*-меры исследуемых алгоритмов для тестовой выборки (N = 15847, 5060 мужчин и 10787 женщин) приведены в таблице 2.

Одним из способов оценки модели в целом, не привязанной к конкретным метрикам, является показатель AUC — площадь (Area Under Curve) под кривой ошибок (Receiver Operating Characteristic curve, ROC). Кривая ошибок представляет из себя линию от (0,0) до (1,1) в координатах True Positive Rate и False Positive Rate.

Таблица 2

Метрики эффективности алгоритмов для тестовой выборки

	DecisionTree	KNeighbors	XGB	RandomForest	GaussianNB
Правильно классифицированные экземпляры	14621	12899	11987	11047	10982
Accuracy (%)	0.92	0.81	0.76	0.7	0.69
Precision, женщины	0.97	0.84	0.75	0.69	0.69
Precision, мужчины	0.84	0.75	0.77	0.81	0.73
Recall, женщины	0.91	0.90	0.95	0.99	0.99
Recall, мужчины	0.94	0.62	0.34	0.07	0.06
f1-score, женщины	0.94	0.87	0.84	0.82	0.81
f1-score, мужчины	0.89	0.68	0.47	0.12	0.11

Составлена авторами на основе собственного исследования

AUC, вычисляя площадь под кривой ROC, указывает, будет ли классификатор с большей вероятностью распределять оценку как положительную, а не как случайно выбранный отрицательный образец. На рисунке 2 представлены кривые ROC всех классификаторов, где хорошо видно, что лучшие модели имеют более высокие значения AUC.

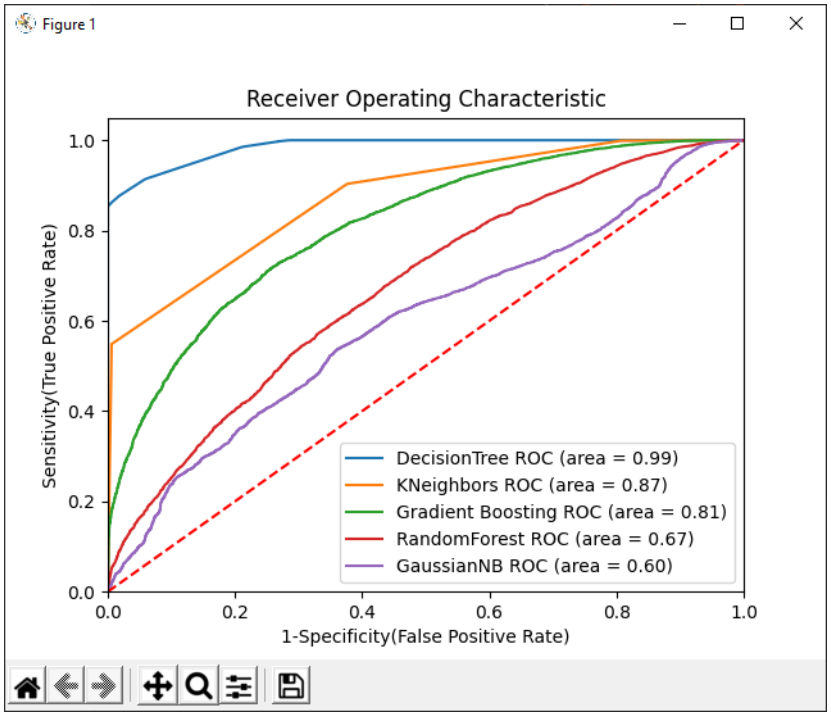

Рисунок 2. *Кривые ошибок для пяти алгоритмов (рисунок авторов)*

Приведенные результаты показывают, что классификатор дерева решений имеет лучшую производительность (AUC = 0,99). Идея построения модели деревьев решений (decision tree) для задачи классификации состоит в том, чтобы осуществлять процесс деления обучающих данных на группы до тех пор, пока не будет достигнут конечный узел дерева, являющийся узлом решения. Совокупность правил, которые дают такое разбиение, позволяя затем делать прогноз для новых данных, — это и есть модель. Графически её можно представить в виде древовидной структуры, где моменты принятия решений соответствуют узлам, имеющим два потомка. В ходе построения дерева решений алгоритм решает основную проблему, с которой связан соответствующий шаг процесса обучения — выбор атрибута, по которому будет производиться разбиение в данном узле. Самый информативный атрибут находится в корне дерева и наилучшим образом разделяет набор данных на классы 1/0.

Структура дерева решений в нашем исследовании (рис. 3) показывает, что в корне дерева — атрибут colorname. Именно он наилучшим образом разделяет набор данных на наши классы.

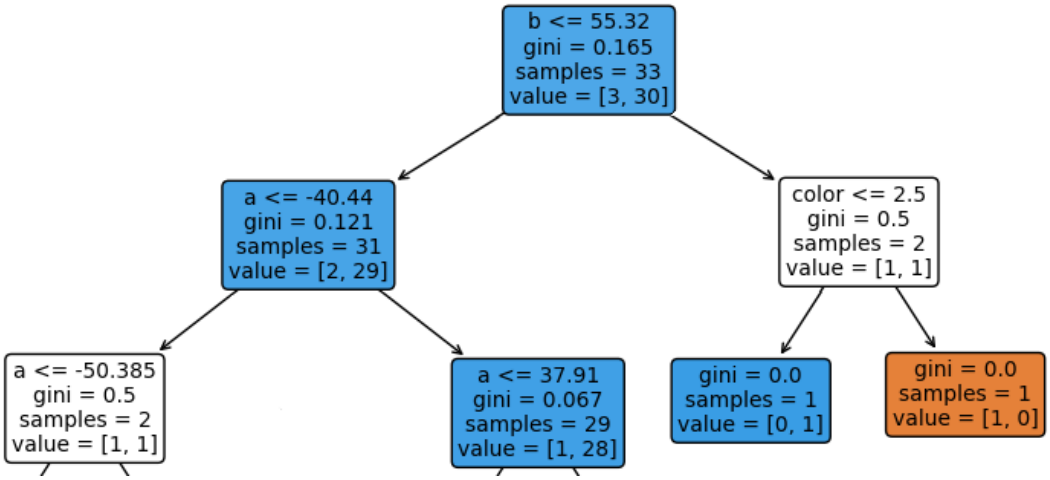

Рисунок 3. *Фрагмент дерева решений (рисунок авторов)*

Эффективность классификатора дерева решений при прогнозировании пола можно оценить на основе анализа матрицы неточностей (confusion matrix), сравнивая фактические и прогнозируемые классы. Содержимое верхней строки этой матрицы — значения TN, FP, значения нижней строки — FN, TP. Таким образом, ячейки матрицы — это количество прогнозов, сделанных алгоритмом дерева решений для тестовой выборки. В нашем исследовании (рис. 4) большинство предсказаний приходится на диагональную линию матрицы, а значит, является правильными предсказаниями.

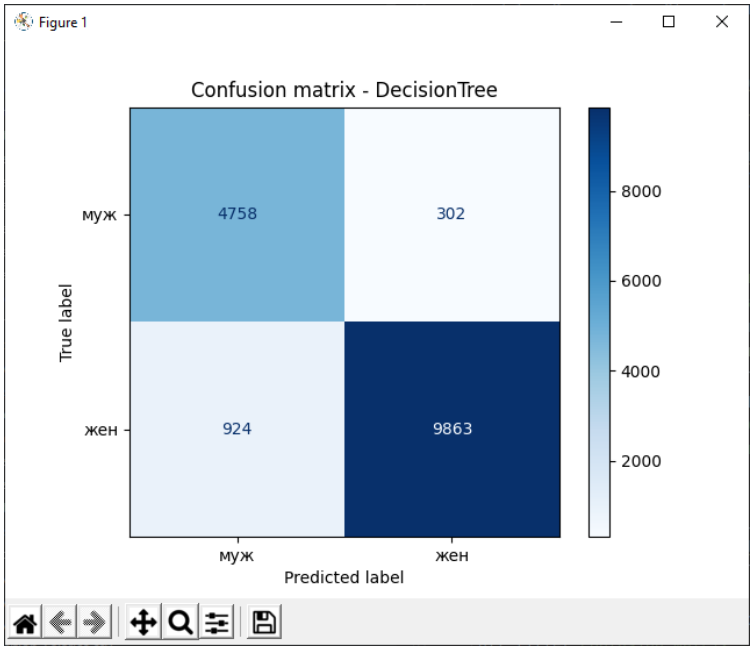

Рисунок 4. *Матрица неточностей дерева решений (рисунок авторов)*

Выводы

В этой исследовательской работе мы проанализировали производительность пяти различных алгоритмов классификации и оценили возможности их применения для прогнозирования пола респондента по его ответам в онлайн-эксперименте, направленном на

изучение социальной дифференциации системы цветонаименований русского языка. Результаты моделирования показывают, что алгоритм дерева решений классифицирует данные с точностью 92 % и качеством, соответствующим значению AUC = 0,99. Это значит, что именно его лучше всего использовать в обработке и анализе полученных данных. Представленная в работе методика оценки эффективности алгоритма классификации с использованием комплекса взаимодополняющих метрик может использоваться в качестве модели для выбора наиболее подходящего программного средства с учетом специфики конкретного случая в дальнейших социологических исследованиях.

ЛИТЕРАТУРА

1. Gibson E., Futrell R., Jara-Ettinger J. [et al.]. Color naming across languages reflects color use // Proceedings of the National Academy of Sciences. — 2017. — № 114(40). — P. 10785–10790. — DOI: https://doi.org/10.1073/pnas.1619666114.
2. Jonauskaite D., Sutton A., Cristianini N. [et al.]. English colour terms carry gender and valence biases: A corpus study using word embeddings // PLoS ONE. — 2021. — № 16(6). — e0251559. — DOI: https://doi.org/10.1371/journal.pone.0251559.
3. Torres A., Serra J., Llopis J., Delcampo A. Color preference cool versus warm in nursing homes depends on the expected activity for interior spaces // Frontiers of Architectural Research. — 2020. — № 9(4). — P. 739–750. — DOI: https://doi.org/10.1016/j.foar.2020.06.002.
4. Jonauskaite D., Abu-Akel A., Dael N. [et al.]. Universal Patterns in Color-Emotion Associations are Further Shaped by Linguistic and Geographic Proximity // Psychological Science. — 2020. — № 31(10). — P. 1245–1260. — DOI: https://doi.org/10.1177/0956797620948810.
5. Грибер Ю.А., Самойлова Т.А., Двойнев В.В. Цветовые предпочтения пожилых людей в различных типах жилого интерьера // Урбанистика. — 2018. — № 4. — С. 36–49. — DOI: https://doi.org/10.7256/2310-8673.2018.4.28349.
6. Konno T., Kakiyama K., Kawabata Y. Observed changes in garment color selection of university students across normal and test periods // AIC2021 Proceedings. — Milano: International Color Association, 2021. — P. 801–806. — URL: https://aic-color.org/resources/Documents/Proceedings_AIC2021_r10.pdf.
7. Грибер Ю.А., Майна Г. Градостроительная живопись. — 2-е изд., испр. и доп. — М.: Издательство Юрайт, 2020. — 104 с. — URL: https://urait.ru/bcode/455980.
8. Николенко С., Архангельская Е., Кадурин А. Глубокое обучение. Погружение в мир нейронных сетей. — СПб: Питер, 2018. — 480 с.
9. Леонов Н.Н. О методике применения машинного обучения в анализе социологических данных // Социологический альманах. — 2019. — Вып. 10. — С. 56–64. — URL: https://socio.bas-net.by/wp-content/uploads/2020/05/Sotsiol_Almanah_Vyp-10.pdf.
10. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. — М: ДМК Пресс, 2017. — 652 с.

11. Griber Y.A., Mylonas D., Paramei G.V. Intergenerational differences in Russian color naming in the globalized era: linguistic analysis // Humanities & Social Sciences Communications. — 2021. — № 8. — P. 262. — DOI: https://doi.org/10.1057/s41599-021-00943-2.

12. Грибер Ю.А. Картография цвета: диагностика развития цветонаименований русского языка с использованием естественно-научных, историографических, социологических и психологических методов. — М.: Согласие, 2021. — 152 с.

13. Paramei G.V., Griber Y.A., Mylonas D. An online color naming experiment in Russian using Munsell color samples // Color Research and Application. — 2018. — № 43. — P. 358–374. — DOI: https://doi.org/10.1002/col.22190.

14. Bonnardel V., Beniwal S., Dubey N. [et al.]. Gender difference in color preference across cultures: an archetypal pattern modulated by a female cultural stereotype // Color Research and Application. — 2018. — № 43. — P. 209–223. — DOI: https://doi.org/10.1002/col.22188.

15. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А.А. Слинкина. — М.: ДМК Пресс, 2015. — 400 с.

16. Груздев А.В. Прогнозное моделирование в IBM SPSS Statistics, R и Python. Метод деревьев решений и случайный лес. — М: ДМК Пресс, 2017. — 652 с.

17. Бринк Х., Ричардс Дж., Феверолф М. Машинное обучение. — СПб: Питер, 2017. — 336 с.

Samoylova Tatyana Arkadyevna
Smolensk State University, Smolensk, Russia
E-mail: tatsamoilova24@gmail.com
RSCI: https://elibrary.ru/author_profile.asp?id=100995

Griber Yulia Alexandrovna
Smolensk State University, Smolensk, Russia
E-mail: y.griber@gmail.com
ORCID: https://orcid.org/0000-0002-2603-5928
RSCI: https://elibrary.ru/author_profile.asp?id=303167
Researcher ID: https://www.researcherid.com/rid/AAG-4410-2019
SCOPUS: https://www.scopus.com/authid/detail.url?authorId=56809444600

A comparative study of the effectiveness of data classification algorithms in sociological analysis of color names

Abstract. The article presents results of a comparative study of different classification algorithms for predicting the gender of the respondent based on his answers in an online experiment aimed at studying the social differentiation of the Russian color-naming system. The data were conducted in an online experiment (http://colournaming.com) in which 2457 native Russian speakers (1402 women, 1055 men), belonging to different age groups ranging from 16 to 98 years old (mean age = 41.36 years, SD = 17.71), participated in 2018–2020. Each of the answers received in the course of the study (N = 55515) contained a number of characteristics of a fundamentally different nature, fixing not only the coordinates of the color samples in the CIELAB system and the color names assigned to them (simple or compound word, phrase or the whole sentence), but also socio-demographic information about the sex and age of the respondent, his place of birth and permanent residence, educational level and profession. The authors analyze various classification algorithms using the software libraries NumPy, Pandas, Scikit-learn for the Python programming language. The effectiveness of the classifiers is evaluated by such parameters as accuracy, precision, F1-score and receiver operating characteristic curve. Simulation results show that the decision tree algorithm classifies data with 92 % accuracy and quality corresponding to AUC (Area Under Curve) = 0.99. This means that it is the best one to use in the processing and analysis of the obtained data. The presented in the paper methodology for assessing the effectiveness of the classification algorithm using a set of complementary metrics can be used as a model for selecting the most appropriate software tool, taking into account the specifics of a particular case in further sociological research.

Keywords: experiment; sociological data analysis; social differentiation of language; color naming; machine learning; classification; Phyton