УДК 004.852

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АЛГОРИТМОВ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ

С.У. Турлакова, Д.А. Ершов

Описан алгоритм обработки и классификации данных по общеобразовательным организацям Челябинской области на основе сведений из форм федерального статистического наблюдения. Исследовано влияние каждого из признаков набора данных на оценки точности классификаторов.

Ключевые слова: методы классификации, деревья решений, случайный лес, автоматическая обработка данных.

Анализ состояния системы образования основывается на результатах обработки форм федерального статистического наблюдения. Форма федерального статистического наблюдения является формуляром – образцом статистического документа, предназначенным для получения в установленном порядке первичных статистических данных, содержащим: вопросы программы наблюдения, место для ответов на них, реквизиты подписи должностного лица, ответственного за предоставление статистической информации, позволяющим осуществлять унификацию процессов сбора и автоматизированной обработки статистической информации [1].

Представление данной формы индивидуально у каждого вида организации: у организаций дошкольного образования, среднего и общего образования, дополнительного образования, среднего профессионального, высшего образования формы кардинально различаются по своему наполнению и объему. Отсутствие инструмента для оперативного и качественного анализа неоднородных массивов данных по образовательным организациям привело к необходимости разработки программы автоматизации извлечения и классификации данных.

На сегодняшний день общеобразовательные организации ежегодно предоставляют сведения по формам статистического наблюдения ОО-1 [2] и ОО-2 [3]. Форма ОО-1 отражает общие сведения об образовательной организации, ее обучающихся и персонале. Форма ОО-2 отражает сведения об имуществе, информационной базе и материально-техническом оснащении образовательной организации. Данные из этих форм являются достоверными и обеспечивают информационную открытость системы образования [4].

На основании статистических данных осуществляется мониторинг качества системы образования, оценка эффективности деятельности

органов местного самоуправления городских округов и муниципальных районов, и в том числе расчет доли муниципальных общеобразовательных учреждений, соответствующих современным требованиям обучения, в общем количестве муниципальных общеобразовательных учреждений [5, 6].

Расчет вышеуказанной доли в рамках оценки эффективности деятельности органов местного самоуправления необходим для обоснования решений о направлении финансовых и трудовых потоков в муниципальные образования области.

Доля рассчитывается на основании 16 показателей из статистической формы ОО-2. Рассматриваются такие показатели, как:

1) наличие водопровода;
2) наличие канализации;
3) наличие отопления;
4) удовлетворительное состояние зданий (то есть не требующих капитального ремонта);
5) отсутствие зданий в аварийном состоянии;
6) наличие спортивного зала;
7) наличие столовой;
8) наличие актового зала или лекционной аудитории;
9) наличие в организации интернета;
10) наличие веб-сайта организации;
11) наличие технических средств для организации дистанционного обучения;
12) обеспеченность пожарной сигнализацией;
13) обеспеченность дымовыми извещателями;
14) обеспеченность пожарными кранами;
15) обеспеченность доступом для маломобильных обучающихся (инвалидов);
16) наличие библиотеки.

Наличие всех вышеописанных признаков (а также отсутствие признаков 4 и 5) является показателем того, что организация соответствует современным требованиям обучения.

Для решения поставленной задачи был разработан следующий алгоритм:

1) Принять на вход некоторое количество папок, каждая из которых соответствует муниципальному образованию области, количество файлов в папке зависит от количества общеобразовательных организаций;

2) Сформировать из принятых папок множество путей элементами которого являются абсолютные пути до папок с исходными файлами,

$$P = \{P_1, P_2, \dots, P_{43}\},$$

при этом P_i, в свою очередь, являются множествами, содержащими абсолютные пути до файлов;

3) Запустить цикл по количеству элементов множества P, внутри которого запустить цикл по длине множества P_i, в котором происходит загрузка файлов в память;

4) Сформировать множество листов

$$S = \{S_1, S_2, \dots, S_{43}\},$$

элементами которого являются множества S_i, содержащие объекты-листы книг исходных файлов.

Формирование набора для классификации осуществляется циклически путем извлечения значений из ячеек в листах множеств S_i. Для обеспечения обезличенности данных организации не разделяются по муниципальным образованиям, а формируют общий список. Результатом работы цикла являются векторы

$$X_i = (x_1, x_2, \dots, x_k),$$

где $i = \overline{1,807}$ – организация (наименования закодированы порядком),
$k = \overline{1,16}$ _ признаки, описанные выше.

Самыми важными признаками являются 4 и 5 (аварийное состояние зданий и необходимость капитального ремонта). Если хотя бы один из данных признаков равен 1, то проход по вектору прекращается и классификатор определяет организацию как несоответствующую современным условиям обучения.

Формально можно записать задачу следующим образом: подобрать такие пороговые значения признаков x_{iu}, где i – номер признака, u – метка, обозначающая задание порога пользователем, при которых оценки ошибок классификаторов accuracy_score, best_score, score не опускаются ниже 90%.

Для обучения классификаторов первичная классификация осуществлялась путем ввода значений из интервалов для каждого признака, впоследствии они подставлялись в следующую логическую конструкцию (рис.1).

$$(x_1 > 0) \wedge (x_2 > 0) \wedge (x_3 > 0) \wedge (x_4 = 0) \wedge (x_5 = 0) \wedge (x_6 \le x_{6u}) \wedge (x_7 \le x_{7u}) \wedge (x_8 \le x_{8u}) \wedge (x_9 \le x_{9u}) \wedge (x_{10} \le x_{10u}) \wedge (x_{11} = 1) \wedge (x_{12} > 0) \wedge (x_{13} \le x_{13u}) \wedge (x_{14} \le x_{14u}) \wedge (x_{15} \le x_{15u}) \wedge (x_{16} \le x_{16u})$$

Рис. 1. Логическая конструкция для первичной классификации

Здесь $x_{6u}, x_{7u}, x_{8u}, x_{9u}, x_{10u}, x_{13u}, x_{14u}, x_{15u}, x_{16u}$ – переменные, значения которых выбираются пользователем и представляют собой пороговые значения признаков в соответствии с их интервалом.

После выполнения описанной процедуры в наборе появляется дополнительный столбец, отвечающий за первичные метки классификации.

Для классификации объектов были использованы деревья решений и метод случайного леса [7–12]. При построении как дерева, так и леса были подобраны такие параметры, как глубина дерева, максимальное количество признаков, по которым ищется лучшее разбиение в дереве, минимальное количество элементов в узле, после которого дерево начинает разделяться. У леса в качестве параметра также выступает количество деревьев решения.

Исходная выборка была разделена на обучающую и тестовую в соотношении 70 на 30 и определены параметры дерева решений. Максимальная глубина дерева задавалась в пределах от 1 до 10, максимальное количество признаков от 1 до 9 (так как после удаления лишних столбцов в наборе осталось 9 столбцов), минимальное количество элементов в узле от 1 до 9. Аналогичным образом выбирались параметры для случайного леса. Количество деревьев в ансамбле – 100.

Также и к лесу, и к дереву была применена кросс-валидация. При применении кросс–валидации модель обучалась K раз на (K-1) подвыборках исходной выборки, а проверялась на одной, но каждый раз на разной. Вследствие этого было получено K оценок качества модели, давших среднюю оценку качества классификации на кросс–валидации. Применение кросс–валидации позволило рассматривать две метрики ошибок: долю правильных ответов и соответствующую среднюю долю правильных ответов на кросс–валидации (табл. 1).

Таблица 1

Сравнение наилучших параметров дерева решений и случайного леса

Наилучшие параметры	**Дерево решений**	**Случайный лес**
max_depth	8	7
max_features	7	8
min_samples_leaf	1	1

Совпадает лишь параметр, отвечающий за минимальное количество признаков в листе, при этом глубина и максимальное количество рассматриваемых признаков для разбиения практически равны.

На рис.2 представлены оценки ошибок построенного дерева решений и случайного леса при введенных пороговых значениях.

```
Введите порог классификации по признаку water от 0 до 7
3
Введите порог классификации по признаку sewerage от 0 до 7
2
Введите порог классификации по признаку heating от 0 до 7
3
Введите порог классификации по признаку internetspeed от 0.0 до 8.0
5
Введите порог классификации по признаку computerswithinternetaccess от 0 до 468
100
Введите порог классификации по признаку firealarm от 0 до 7
2
Введите порог классификации по признаку smokedetectors от 0 до 7
2
Введите порог классификации по признаку firecranes от 0 до 5
2
Введите порог классификации по признаку fordisabled от 0 до 4
1
```

Рис. 2. Параметры логической конструкции, на которой производилось первичное обучение классификаторов

Значения ошибок для дерева решений и случайного леса представлены в табл. 2.

Таблица 2

Сравнение ошибок дерева решений и случайного леса

Оценка ошибки	**Дерево решений**	**Случайный лес**
accuracy_score	1.0	0.9851485148514851
best_score	0.9900826446280991	0.9834710743801652

По данным из таблицы наблюдается незначительное преимущество дерева, однако в целом оба классификатора показывают очень высокие результаты по точности.

Для сравнения исходная задача классификации была решена с применением наивного байесовского классификатора. Лучшие параметры обучения представлены в табл. 3.

Таблица 3

Параметры обучения наивного байесовского классификатора

Параметр	**Значение**
alpha	1.0
binarize	0.0
class_prior	None
fit_prior	True

Параметр alpha является логическим и описывает применение аддитивного параметра сглаживания (1 означает, что сглаживание применено). Параметр binarize означает параметр бинаризации образцов объектов. В данном случае он равен 0, так как входные данных уже частично состоят из двоичных векторов. Class_prior определяет корректирование априорных вероятностей в соответствии с данными и также является логическим. Так как его значение None, априорные вероятности не корректируются. Fit_prior определяет, необходимо ли классификатору изучать класс предшествующих вероятностей (в случае с рассматриваемым набором данных классификатор определил необходимость их рассмотрения).

Наивный байесовский классификатор показал точность 0.912396694214876.

Было произведено варьирование всех признаков, чтобы определить, какие из них оказывают наибольшее влияние на точность классификации. Выяснилось, что признаком, изменение которого сильнее всего влияет на точность, стал признак computerswithinternetaccess, однако и его изменение не колебало точность классификаторов более, чем на 5%.

Полученные результаты оказались достаточно точны, но при этом не стоит забывать, что параметры логической конструкции классификации подбирались случайным образом. После проведения вычислительных экспериментов можно сделать вывод о том, что изменение 8 из 9 численных признаков практически не влияет на точность оценки, однако изменение параметра internetspeed значительно повлияло на классификаторы, понизив их точность. Также стоит отметить, что наиболее чувствительным к изменению признаков оказался наивный байесовский классификатор, оценка ошибки которого варьировались от 89,6% до 92,07%, что хуже оценок дерева и леса.

Заключение. Предложенная методика автоматической обработки данных позволяет оперативно решать поставленные задачи, возникающие в системе образования. Была разработана программа, которая позволила автоматически обрабатывать формы федерального статистического наблюдения, после чего полученный набор можно было применять для обучения классификаторов. Так как все классификаторы являются методами обучения с учителем, был реализован инструмент первичной классификации на основе значений, вводимых пользователем, в интервалах столбцов из полученного набора.

Также было исследовано влияние каждого из признаков набора на оценки точности классификаторов. Установлено, что подавляющее большинство признаков незначительно изменяет оценки ошибок, однако при этом варьирование параметра internetspeed значительнее всего влияет на точность классификации. Стоит отметить, что из рассматриваемых

классификаторов самым чувствительным к изменению признаков оказался наивный байесовский классификатор.

По данным измерительного контроля были построены гистограммы и определены законы распределения случайных величин параметров. Интегрированием функций распределения, ограниченных пределами допусков, были получены уровни дефектности (частота дефекта).

Библиографический список

1. Федеральный закон от 29.11.2007 N 282-ФЗ (ред. от 18.04.2018) "Об официальном статистическом учете и системе государственной статистики в Российской Федерации" [Электронный ресурс]// http://www.consultant.ru/document/cons_doc_LAW_72844/2680848101fddc00e6720c aa1e6d53fdff937c72/ – Дата доступа: 07.05.2020

2. Приказ Росстата от 12.08.2019 N 441 (ред. от 30.08.2019) "Об утверждении формы федерального статистического наблюдения с указаниями по ее заполнению для организации Министерством просвещения Российской Федерации федерального статистического наблюдения в сфере общего образования" [Электронный ресурс] // http://www.consultant.ru/document/cons_doc_LAW_331760/2ff7a8c72de3994f30496a 0ccbb1ddafdaddf518/ – Дата доступа: 07.05.2020

3. Приказ Росстата от 01.11.2019 N 648 (ред. от 05.12.2019) "Об утверждении форм федерального статистического наблюдения с указаниями по их заполнению для организации Министерством просвещения Российской Федерации федерального статистического наблюдения в сфере общего и среднего профессионального образования [Электронный ресурс] // http://www.consultant.ru/document/cons_doc_LAW_337009/ (04.05.2020)

4. Федеральный закон от 29.12.2012 г. № 273-ФЗ «Об образовании в Российской Федерации» [Электронный ресурс] // http://www.consultant.ru/document/cons_doc_LAW_140174/ – Дата доступа: 08.05.2020.

5. Об осуществлении мониторинга системы образования: постановление Правительства РФ от 05.08.2013 № 662 (ред. от 25.05.2019) [Электронный ресурс] // http://base.garant.ru/70429494/ – Дата доступа: 07.05.2020.

6. Постановление Правительства РФ от 17.12.2012 N 1317 (ред. от 16.08.2018) "О мерах по реализации Указа Президента Российской Федерации от 28 апреля 2008 г. N 607 "Об оценке эффективности деятельности органов местного самоуправления городских округов и муниципальных районов" и подпункта "и" пункта 2 Указа Президента Российской Федерации от 7 мая 2012 г. N 601 "Об основных направлениях совершенствования системы государственного управления" [Электронный ресурс] // http://www.consultant.ru/document/cons_doc_LAW_139508/ – Дата доступа 03.05.2020.

7. Шахиди, А. Деревья решений – общие принципы работы [Электронный ресурс] // https://loginom.ru/blog/decision-tree-p1 (Дата доступа 21.04.2020)

8. Кафтанников, И.Л. Особенности применения деревьев решений в задачах классификации / И.Л. Кафтанников, А.В. Парасич // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2015. – Т. 15, №3.– С. 26–32. DOI:10.14529/ctcr150304

9. Breiman, L. Random Forests / L. Breiman // Machine Learning : journal. – 2001. – Vol. 45, no. 1. – P. 5–32. – DOI:10.1023/A:1010933404324

10. Breiman, L. Bagging Predictors / L. Breiman // Machine Learning, 24: pp. 123–140. 11 Мюллер, А. Введение в машинное обучение с помощью Python / А. Мюллер, С. Гвидо – Вильямс, 2017 – 480 с.

11. Мюллер, А. Введение в машинное обучение с помощью Python / А. Мюллер, С. Гвидо – Вильямс, 2017 – 480 с.

12. Domingos, P On the optimality of the simple Bayesian classifier under zero-one loss/ P. Domingos, M. Pazzani // Machine Learning, 29:103–137.