УДК 519

ПРИМЕНЕНИЕ ДЕРЕВЬЕВ РЕШЕНИЙ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ФОРМЫ ПОПЕРЕЧНОГО ПРОФИЛЯ ГОРЯЧЕГО ПРОКАТА

А.Н. Шкарин, магистрант гр. М-ПМ-15, I курс
Научный руководитель: Е.В. Кузнецова, к.ф.-м.н., доцент
Липецкий государственный технический университет
г. Липецк

Под поперечным профилем горячего проката в данной статье будет иметься в виду поперечное сечение полосы (листа) характеризуемое разными толщинами в центре и у краев. Исходными данными является массив измерения толщины 1500 полос по ширине с шагом измерения 5 мм, усредненный по длине. Таким образом, профиль горячего проката – некоторая дискретная функция, которая быстро и на большие величины изменяется вблизи кромок.

Традиционный подход для моделирования зависимости толщины полосы $h(x)$ от удаления x от середины – применение полиномиальной регрессии. Данная зависимость имеет следующий вид

$$h(x) = a + a_1 x^2 + \ldots + a_n x^{2n}, \quad (1)$$

где a – толщина на середине полосы; $a_1 \ldots a_n$ – параметры профиля.

Согласно источнику [1], при степени $n > 6$ начинает проявляться известное свойство полиномов больших порядков – качество аппроксимации ухудшается. Тогда, воспользовавшись формулой (1), получаем модель следующего вида

$$h(x) = a + a_1 x + a_2 x^2 + a_3 x^4 + a_4 \mathrm{x}^6, \quad (2)$$

где параметр a_1 учитывает асимметрию.

Для того, чтобы исключить влияние толщины профиля на исходный результат, решено было пронормировать выборку от нуля до единицы по следующей формуле:

$$h(x_i) = \frac{h_i - h_{\min}}{h_{\max} - h_{\min}},$$

где $h_{\min}$ – минимальное значение исходного профиля; $h_{\max}$ – максимальное.

Коэффициенты модели (2) найдем с помощью МНК [2], смысл которого состоит в минимизации квадратов отклонений экспериментальных данных от данных, полученных от найденной модели

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (h(x_i) - \hat{h}(x_i))^2 \longrightarrow \min,$$

где $h(x_i)$ – найденные значения, $\hat{h}(x_i)$ – экспериментальные данные.

В ходе работы с выборкой было выяснено, что существует четыре основных вида профилей, показанных на рисунке 1.

Пусть $\{x^{(i)}\}$, $x^{(i)} = [a, a_1, ...a_4]^T$ – множество векторов признаков; $C = \{C_j\}$, $j = 1,...,4$ – множество непересекающихся классов; $\{c^{(i)}\}$, $i = 1,...,m$ – множество классов для известных объектов; множество пар $\{x^{(i)}, c^{(i)}\}$, $i = 1,...,m$ – обучающее множество.

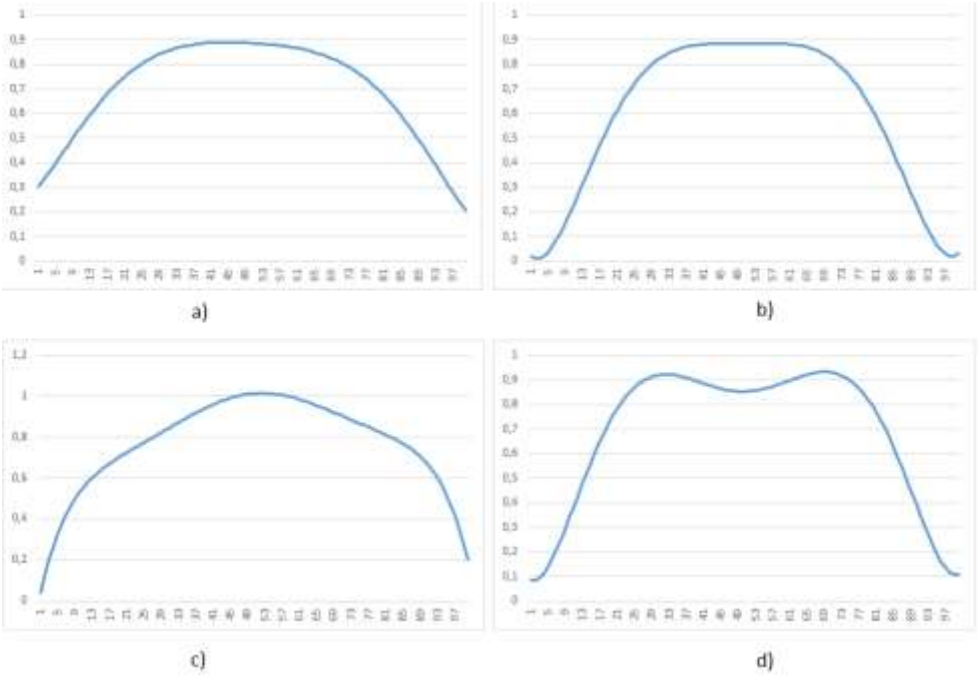

Рисунок 1. Основные виды профилей горячего проката

Поскольку существует конечное множество классов, имеющих одни и те же признаки, и существует обучающая выборка, имеет смысл решать задачу классификации [3]. Основная цель – нахождение функции $f : X \mapsto C$.

Для проверки эффективности полученной далее модели введем понятие матрицы классификации (матрицы потерь). Это квадратная матрица, размер которой зависит от количества рассматриваемых классов. Столбцы этой матрицы резервируются за экспертными решениями, а строки за решениями классификатора. При классификации мы инкрементируем число, стоящее на пересечении строки класса, который вернул классификатор и столбца класса к которому действительно относится документ.

$$Confusion\ Matrix = \begin{bmatrix} Известные \,|\, Предсказанные & C_1 & C_2 & ...C_4 \\ C_1 & t_{11} & e_{12} & ...e_{14} \\ ... & ... & ... & ... ... \\ C_4 & e_{41} & e_{42} & t_{44} \end{bmatrix} \quad (3)$$

Диагональные элементы – правильные ответы, вне диагональные – ошибки. Для количественной оценки качества классификации вычислим точность по формуле

$$Acc = \frac{\sum_{i=1}^{4} t_{ii}}{m}, \quad (4)$$

где m – общее количество элементов обучающего множества.

Для оценки адекватности полученной модели воспользуемся принципом кросс-проверки. Разобьем имеющуюся выборку на тестовую (20%) и обучающую (80%). Если результаты на тестовом и обучающем множестве приблизительно равны, то цель моделирования достигнута.

В качестве метода классификации применили деревья решений, которые обладают следующими преимуществами: интуитивно понятны, интерпретируемы, эффективны [4]. Общий алгоритм решения задачи классификации можно представить следующим образом:

1. Сбор данных
2. Подготовка
3. Анализ (визуальная проверка дерева)
4. Обучение
5. Проверка адекватности

Базовый вариант дерева решений – представление правил с помощью иерархической, последовательной структуры бинарного дерева. Для определения, к какому классу принадлежит объект, необходимо двигаться от корневой вершины до листа. В каждой вершине происходит развилка. Каждый лист – класс C_j.

Основной принцип построения такого дерева – разделяй и властвуй. Каждое обучающее множество рекурсивно разбивается на подмножества, в каждом из который доминирует (властвует) один из классов.

Общий алгоритм построения дерева решений:

1. Выбирается переменная, которая помещается в корень дерева
2. Из вершины строятся ветви, соответствующие всем возможным значениям выбранной независимой переменной
3. Множество объектов из обучающей выборки разбивается на несколько подмножеств в соответствии со значением выбранной переменной. В каждом подмножестве будут находится объекты, у которых значение выбранной переменной одно и то же.

Построим дерево решений для данной выборки с помощью алгоритма C4.5 [5], который реализован в пакете RWeka языка R, рисунок 2.

Тогда матрицы классификации для обучающего $Confusion\ Matrix_{train}$ и тестового $Confusion\ Matrix_{test}$ множества, найденные по формуле (4) имеют вид

$$Confusion\ Matrix_{train} = \begin{matrix} 169 & 1 & 2 & 1 \\ 3 & 73 & 0 & 0 \\ 3 & 0 & 65 & 0 \\ 0 & 0 & 0 & 99 \end{matrix} \quad Confusion\ Matrix_{test} = \begin{matrix} 42 & 1 & 0 & 1 \\ 0 & 16 & 0 & 0 \\ 1 & 0 & 16 & 0 \\ 0 & 1 & 0 & 23 \end{matrix}$$

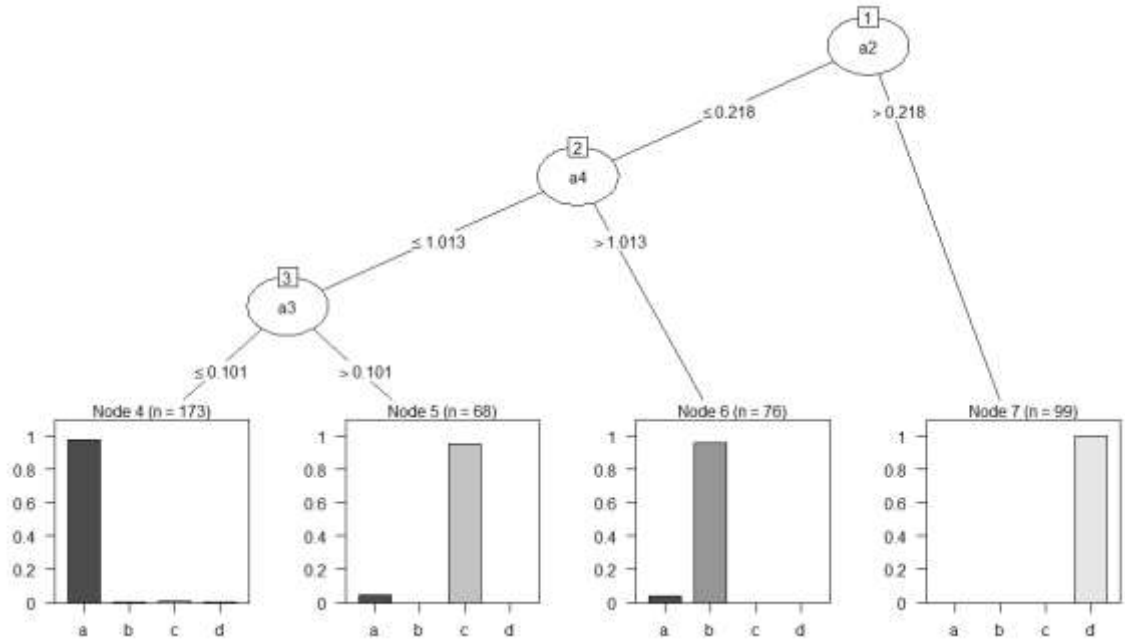

Рисунок 2. Графическое представление полученного дерева решений

Точность классификации, найденная по формуле (4) соответственно равна $Acc_{train} \approx 0.95$ и $Acc_{test} \approx 0.92$. Следовательно, подобрана максимально достоверная модель.

Список литературы:

1. Робертс, В.Л. Холодная прокатка стали [Текст] / П.И. Полухин. – М.: Металлургия, 1982. – 544 с.

2. Айвазян, С.А. Прикладная статистика. Основы эконометрики: Учебник для вузов: В 2 т. [Текст] / С.А. Айвазян. – М.: ЮНИТИ-ДАНА, 2001. – 432 с.

3. Фор, А. Восприятие и распознавание образов [Текст] / А.В. Серединский. – М.: Машиностроение, 1989. – 292 с.

4. Hastie, T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. 2009.

5. Технологии анализа данных [Сайт]. URL: http://www.basegroup.ru