УДК 519.2, 004.8

Применение методов data mining с целью выявления зависимости температуры заготовки от истории её нагрева.

Жуков П.И., Глущенко А.И.

Старооскольский технологический институт им. А.А. Угарова (филиал) НИТУ «МИСиС»

Application of data mining methods to identify the dependence of the temperature of the billets on the history of its heating.

Glushchenko A.I, Zhukov P.I.

Stary Oskol technological institute n.a. A.A. Ugarov (branch) NUST "MISIS"

Аннотация:

В статье рассмотрен процесс применения протокола статистических исследований и концепций Data Mining к технологическим данным. Построена модель зависимости температуры заготовки от истории её нагрева с конечной точностью прогноза ~ 10 ºС

Ключевые слова: Data Mining**,** разведочный анализ, деревья решений, градиентный бустинг, модели смешенных взаимодействий.

Abstract:

The article discusses the process of applying the protocol of statistical analysis and Data Mining concepts to technological data. A model of the dependence of the temperature of the billets on the history of its heating with a finite forecast accuracy of ~ 10 ºС

Keywords: Data Mining, exploration analysis, decision trees, gradient boosting, mixed-effects models.

Введение.

Постоянно растущие требования к объемам и качеству выпускаемой продукции приводят к тому, что технологические процессы (ТП) становятся более наукоемкие и ответственные, а технологические объекты (ТО) более сложными и комплексными. Вместе с ростом требований к ТО растут и требования к АСУ ТП.

Характерной особенностью современных систем управления в сложившейся ситуации является то, что в процессе работы они собирают, архивируют и сохраняют большие объемы полезных данных, которые, как правило, используются только для фиксации нормального технологического режима работы.

Одновременно с этим, в связи с развитием вычислительной техники, стали приобретать популярность методы машинного анализа больших объемов, в том числе и технологических данных (Data Mining) с целью их формализации и полезного применения (Business Intelligence - BI) [1]. Было принято решение проанализировать эти данные с целью выявить зависимости, которые можно было бы полезно использовать в дальнейшем.

Теоретические предпосылки. Сбор и предобработка данных.

В качестве исследуемого объекта предлагается рассмотреть АСУ ТП печей нагрева СПЦ-1 АО ОЭМК. Выбор данного технологического объекта в составе конкретного предприятия в том числе обусловлен четкой необходимостью в моделях, способных описать зависимость между температурой заготовки на стане и температурами внутри печи [2]. Предполагается, что построение подобной модели позволит повысить качество конечного продукта и эффективность конкретного участка технологической цепи [3].

Печи нагрева СПЦ-1 АО ОЭМК представляют собой методические шести-зонные пламенные печи по технологическому назначению разделенные на три пары зон: 1) Зоны №1 и №2 – методические; 2) Зоны №3-4 – сварочные; 3) Зоны №5-6 – томильные. В каждой паре зон установлены термоэлектрические преобразователи (термопары) для контроля и регулирования режима нагрева. Выгрузив из АСУ ТП информацию с термопар, можно получить временной ряд температур внутри печи.

В настоящий момент АСУ ТП печей нагрева СПЦ-1 АО ОЭМК не ведет позаготовочное слежение и фиксирует только посад и выдачу заготовки. Однако, известно, что транспортировка непрерывно литых заготовок (НЛЗ) через печь осуществляется при помощи шагающего механизма, количество «шагов» которого в каждой из зон строго регламентировано: 1) Зона №1 и №2 – с 1-го по 35-й шаг; 2) Зона №3 и №4 – с 36-го по 58-й шаг; 3) Зона №5-6 – с 59-го по 68-й шаг. Всего заготовка делает 68 шагов с временным интервалом не менее 2-х минут. При условии, что посад новой заготовки сопровождается выгрузкой старой (всего в печи 68 позиций), выгрузив из АСУ ТП такие данные можно получить пространственный набор данных по заготовкам. На основании вышеизложенного было решено смоделировать концептуальную модель первичной обработки данных и организации хранилища (рис.1).

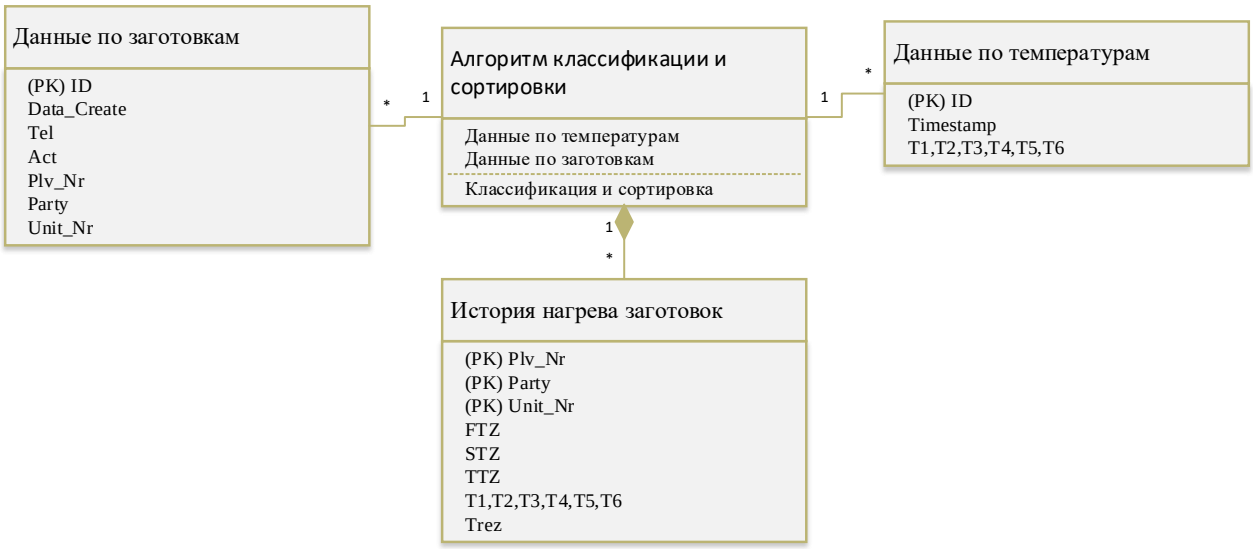

Рис. 1 Концептуальная схема организации хранилища данных

Здесь: 1) ID – первичный ключ в обеих таблицах; 2) Timestamp – временная шкала; 3) T_1, T_2, T_3, T_4, T_5 и T_6 – температуры в соответствующих зонах; 4) Data_create – временные метки заготовки; 5) Tel – номер телеграммы; 6) Act – телеграмма; 7) Plv_Nr – номер плавки; 8)Party – единица прокатки; 9) Unit_Nr – номер заготовки плавки; 10) FTZ, STZ, TTZ – время заготовки в первой, второй и третьей паре зон; 11) Trez – температура заготовки на стане. Из 17175 записей о посаде и выдачи заготовок в печь, была сформирована сводная история нагрева для 8526 заготовок. Предлагается рассмотреть возможность прогнозировать температуру заготовки на стане (Trez) используя историю её нагрева.

Разведочный анализ данных. Построение моделей.

Основной целью применение разведочного анализа (РДА) является «углубление» в данные, с целью обнаружить аномалии и отклонения, проверить основные гипотезы и найти основные закономерности [4]. Единого алгоритма проведения анализа нет, однако в последнее время набирает популярность протоколе статистических исследований (A protocol for data exploration to avoid common statistical problems) [5], в рамках которого определены основные шаги РДА, которым и предлагается следовать.

Результаты оценки выбросов (рис.2) показали, что для всех параметров, том числе и отвечающих за температуры в зонах, характерно большое количество значений выше медианного. Данный факт свидетельствует о необходимости более тщательной настройки моделей и применении средств, повышающих её робастность.

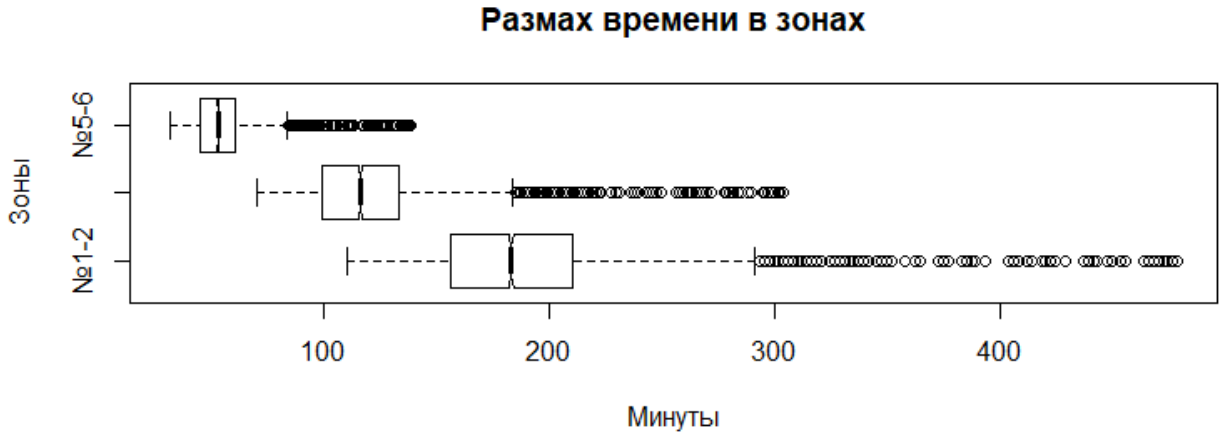

Рис. 2 Оценка выбросов в данных

Проведение формального теста Колмогорова-Смирнова для всей выборки и теста Шапиро-Уилка для равных интервалов показало вероятностное распределение исследуемых параметром отличное от нормального.

При поиске коллинеарных переменных необходимо учесть тот факт, что из контекста предметной области нам известно: 1) Зоны регулируются попарно; 2) Общее время заготовки в печи является суммарным параметром. Для поиска коллинеарных зависимостей использовался фактор инфляции дисперсии (1), связанный с дисперсией невязок и дисперсией

признаков линейной регрессии j-й объясняющей переменной на остальные объясняющие переменные через коэффициент детерминации R_j^2

$$VIF_j = 1/(1-R_j^2) \quad (1)$$

После замены заведомо коллинеарных параметров на их арифметическое среднее, тест VIF не выявил других коллинеарных зависимостей в выборке.

Так как вероятностные распределения параметров выборки отличны от нормальных применим критерий Спирмана (2) для оценки корреляций, а также оценим уровень значимости с использованием t-статистики (3).

$$\rho = 1 - ((6 * \sum d^2) / n * (n^2 - 1)) \quad (2)$$

Где: 1) d – разность каждых значений по рангу; 2) n – размерность выборки. По результатам теста было выяснено: 1) Параметры слабо коррелируют с целевой переменной по отдельности; 2) Все коэффициенты корреляции являются значимыми.

На основании проведенного РДА с учетом все граничных критериев было решено использовать для построение зависимости модели смешенных взаимодействий (Mixed-Effects Models) На основе множественного эксперимента с различными взаимодействиями предикторов была получена следующая модель (4).

$$f(y_j) = \beta_{0j} + \sum_{i=1}^{6}\sum_{k=1}^{p} \beta_{ik} * x_{ij}^k + \sum_{m=1}^{4}\sum_{k=1}^{p} \beta_{mk} * (x_{1j}^m * x_{3j}^{(p+1)-k}) \quad (3)$$

Где: 1) p = 4 – степень используемого полинома; 2) m = 4 – коэффициент, позволяющий определить наивысший порядок эффектов для сочетаний выбранных взаимодействий; 3) $x_{1,\,j}$– это j-е значение параметра FTZ, 4) $x_{3,\,j}$– это j-е значение параметра TTZ (см.табл.8). После апробации модели были оценены следующие показатели: 1) Среднеквадратичная ошибка модели – 744; 2) Средняя ошибка модели в абсолютных значениях – 23 ^{o}C; 3) Средняя ошибка прогноза модели 15 ^{o}C

Для увеличения точности прогноза было принято решение применить композиционный алгоритм и смоделировать ансамбль из моделей (4) в виде дерева решений, где каждая последующая итерация будет учитывать ошибки предыдущей (Stochastic Gradient Boosting). Чтобы полученная модель была более робастной и менее подверженной к переобучению применим регуляризатор вида эластичной сети (5).

$$\hat{\beta} = \arg\min\left(\sum_{j=1}^{n}(y_j - f(y_j))^2 + \lambda_1(\beta)^2 + \lambda_2|\beta|\right) \quad (4)$$

Где: n – размерность выборки; $\lambda_1 = 0.001$ и $\lambda_2 = 0.5$ – штрафующие коэффициенты; $f(y)$ – регрессионная модель (9); β – параметры этой модели. После применения градиентного дерева решений были повторно оценены показатели качества: 1) Среднеквадратичная ошибка – 504; 2) Средняя ошибка модели в абсолютных значениях – 18 ^{o}C; 3) Средняя ошибка прогноза модели 10 ^{o}C.

Заключение

В процессе разведочного анализа было выявлено, что полученные данные содержат большое количество выбросов, в основном выше своего медианного значения. Предполагается, что данный факт не позволяет считать вероятностное распределения параметров нормальным в виду чего невозможно применить более мощные параметрические методы для анализа и построения моделей зависимости. Однако, при этом показатели качества моделей, полученных, в результате экспериментов, удовлетворяют допустимым технологической инструкцией температурным интервалам ~20 ^{o}C. Предполагается, что дальнейшее увеличение точности моделей возможно только при условии уменьшения зашумленности данных, возможное в следствии более точной настройки алгоритмов сортировки и классификации на этапе сбора и предобработки данных.

Список источников.

1. Прокопенко Н.Ю. Применение интеллектуальных информационных систем и методов data mining в промышленности / Н.Ю. Прокопенко, М.С. Прокопенко // Информатика: проблемы, методология, технологии материалы XVIII Международной научно-методической конференции : в 7 т.. Воронежский государственный университет. 2018. С. 318-323.
2. Сердобинцев Ю. П. Перспективные направления повышения качества функционирования технологического оборудования: монография / Ю. П. Сердобинцев, О. В. Бурлаченко, А. Г. Схиртладзе. – Старый Оскол: ООО «Тонкие наукоемкие технологии», – 2010. – 412 с.
3. Андреев С.М. Совершенствование информационного обеспечения энергосберегающих режимов нагрева металла // Машиностроение: сетевой электронный научный журнал. 2015. Т. 3. № 2. С. 3-10.
4. П. Брюс, Э. Брюс. Разведочный анализ данных // Практическая статистика для специалистов Data Science. — СПб.: БХВ-Петербург, 2018.— С. 19—58.
5. Zuur AF, Ieno EN, Elphick CS..A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 1—2010 -- Pp. 3-14
6. Gałecki, Andrzej; Burzykowski, Tomasz Linear Mixed-Effects Models Using R: A Step-by-Step Approach. New York: Springer, – 2013, – 542 p.