Киселева В.А., Рыжкова М.Н.
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
masmash@mail.ru

Обзор методов анализа малой выборки

Хранилище данных – класс систем в IT, в которых данные разной природы (логистика, производство и другое) с точки зрения бизнес-процессов собираются из многих источников для построения различных аналитик. Одна из задач хранилища – это обеспечение качества данных. Информация загружается из систем-источников, в которых могут быть ошибки, наличие которых неважно для производственного процесса, но при построении аналитики важна корректность и размеченность данных. Производственных процессов большое количество, каждый из которых знать невозможно, но нужно масштабно оценивать качество данных. Здесь стоит воспользоваться математическим прогнозированием, на основании которого можно будет делать выводы о качестве данных.

Математически прогнозировать ту или иную величину, например объем выручки или объем производства, следует на основании исторических периодов, которые уже были выверены, в качестве данных которых есть уверенность, вычислять прогноз на следующий период. Это позволит сравнить факт с теорией, и в случае отклонения заранее узнать о наличии потенциальных ошибок.

Следовательно, необходимо разработать систему, которая позволит узнать об ошибке в процессе загрузки данных до составления аналитики по ним. Сложность в том, что исторический ряд короткий, в основном величины характеризуются месяцем, а данные старше года нерелевантны, так как экономика нестабильна. Здесь стоит применять алгоритмы для прогнозирования показателей в условиях малой выборки.

Выборка считается малой в случае, если она содержит недостаточно информации для получения заданной точности и достоверности при решении определенной задачи.

В статье [1] проведен анализ малой выборки экспериментальных данных, позволяющий принимать верные решения по прогнозированию, а также разработаны адекватные модели регрессионные модели, используемые для предсказания. Методика исследования строится на новом методе прогнозирования «метод скользящей матрицы», который заключается в непрерывном обновлении коэффициентов регрессионной модели путём удаления строки с устаревшими данными и ввода новой строки с данными в прогнозируемой точке.

В статье [2] анализируется проблема прогнозирования при малой выборке. Теория и опыт продемонстрировали, что при верном способе организации выборочного наблюдения можно подчерпнуть достоверную информацию об исследуемой совокупности. Выборочный метод дает возможность при малом количестве исследуемых единиц извлечь объективные данные из всей исследуемой совокупности. В статье рассмотрен морфологический анализ как метод прогнозирования, который осуществляется при помощи матрицы характеристик объекта прогнозирования и их вероятных значений с дальнейшим перебором и оценкой вариаций комбинаций рассматриваемых значений.

В работе [3] представлен обзор текущих методов обучения нейронных сетей на основе малых выборок. Small Sample Learning (SSL) – это новая парадигма обучения, которая направлена на моделирование способностей людей к обучению. На данный момент существует 4 подхода при реализации обучения на основе опыта, которые были разработаны в условиях недостаточного объема выборки:

1) увеличение объёма выборки последующая реализация классических методов машинного обучения;
2) использование малых выборок для исправления известных моделей/извлеченных знаний;
3) уменьшение зависимости классических методов машинного обучения от количества выборок для создания алгоритмов, применимых к малым выборкам;

4) мета-обучение.

В статье [4] рассматривается применение многомерного метода точечных распределений для построения статистически значимой математической модели для прогнозирования по исходным данным малого объема. Метод заключается в построении выборки виртуально объёма, которая может быть обработана классическими методами прогнозирования.

В работе [5] рассмотрен метод оценки прогнозируемого значения случайного процесса, отсчеты которого коррелированы. В задачах, когда объем статистических данных ограничен, наиболее применимыми оказываются экстраполяционные методы прогнозирования, к которым относятся метод наименьших квадратов и совокупность методов сглаживания. Здесь проанализирована возможность применения метода наименьших квадратов для прогнозирования в условиях малой выборки, например, случайного процесса спроса, наблюдения которого коррелированы.

С статье [6] рассмотрено применение статистик, использующих размах выборки, к обработке измерений, которые распределены по нормальному закону. При помощи данного способа происходит выявление результатов, содержащих грубые ошибки. Эффективность оценки результатов измерений для малых выборок существенно уступает эффективности среднего арифметического значения, это объясняется тем, что при вычислении полу суммы крайних членов выборки используют всего два результата измерений из n-объема выборки.

Литература

1. Кумаритов, А. М. Анализ малой выборки экспериментальных данных при управлении газоснабжением региона / А. М. Кумаритов, А. Э. Дзагоев, Р. Б. Шарибов. — Текст : непосредственный // Известия высших учебных заведений. Проблемы энергетики. — 2018. — № 1-2. — С. 62-69.

2. Тюлькина, Н. В. Прогнозирование при малой выборке / Н. В. Тюлькина, А. П. Корнеева, А. В. Селиверстова. — Текст : непосредственный // Теория и практика современной науки. — 2018. — № 12(42). — С. 613-619.

3. Shu J., Zongben X., Deyu M. Small sample learning in big data era // arXiv.org, 2018. 76 p. arXiv:1808.04572v3 [cs.LG]

4. Попукайло, В. С. Прогнозирование уровня успеваемости абитуриентов в условиях малой выборки / В. С. Попукайло. — Текст : непосредственный // Наука вчера, сегодня, завтра. — 2016. — № 3 (25) . — С. 95-101.

5. Головко, В. А. Прогнозирование коррелированного временного ряда по малой выборке исходных данных / В. А. Головко, Я. Хазим. — Текст : непосредственный // Вестник НТУ "ХПИ. — 2014. — № 35(1078). — С. 43-47.

6. Гордеев, В. А. Статистические процедуры при обработке малых выборок / В. А. Гордеев, Г. Г. Шевченко. — Текст : непосредственный // Известия высших учебных заведений. Геодезия и аэрофотосъемка. — 2021. — № 2. — С. 152-157.