DOI: 10.34185/1991-7848.itmm.2020.01.035

АНАЛИЗ ОСОБЕННОСТЕЙ ЭТАПА ПОДГОТОВКИ ВЫБОРКИ ДЛЯ ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНОЙ НЕЙРОННОЙ СЕТИ

Захаров А.А., Селиверстова Т.В.

Национальная металлургическая академия Украины

Аннотация. Авторы дают обобщенную характеристику этапу подготовки данных (Data Preparation), используемому в интеллектуальном анализе данных (Data Mining) для обучения нейронных сетей. Выделяются и описываются характерные особенности процесса составления выборки (dataset) из генеральной совокупности данных, извлечения и генерации признаков. Дается классификация выборки, а также обосновывается мысль о том, что классические подходы к этапу подготовки данных в интеллектуальном анализе данных, не применимы при подготовке данных для обучения и использования генеративно-состязательной нейронной сети (generative-adversarial network).

Ключевые слова: НЕЙРОННЫЕ СЕТИ, ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ, ПОДГОТОВКА ДАННЫХ, ВЫБОРКА, ПЕРЕМЕННЫЕ ЛАТЕНТНОГО ПРОСТРАНСТВА, ЗНАЧЕНИЯ ПРИЗНАКОВ.

Подготовка данных или Data Preparation выделяется стандартами CRISP-DM, SEMMA и другими в отдельный процесс. И действительно, подготовка материалов для обучения нейронной сети является решающим фактором в скорости обучения и точности предсказаний сетью в будущем. Согласно различным отчетам практически 80% временных затрат в задачах обучения нейронных сетей приходится именно на подготовку данных.

Для целей обучения нейронных сетей совокупный набор исходных данных принято называть генеральной совокупностью.

Одним из основных этапов подготовки данных для обучения является порождение выборки (datasets) из генеральной совокупности.

Методы формирования выборок зависят от класса решаемой задачи:

- задачи классификации требуют разделения данных в том же соотношении как в генеральной совокупности;

- задачи регрессионного анализа требуют одинакового распределения целевой переменной в различных выборках.

Принято считать, что существует три вида выборок:

- валидационная (validation sample);

- контрольная (test sample);

- обучающая (training sample).

Каждый из этих трех типов выборок необходим на разных этапах обучения нейронной сети. В некоторых случаях валидационная и тестовая выборки могут состоять всего из 10% от обучающей выборки. В классических случаях данные из всех трех выборок не должны пересекаться и зависеть друг от друга. После подготовки данных необходимыми этапами являются очистка данных и генерация признаков (feature generation) или их извлечение (feature extraction).

Если учесть, что признак(feature) является отдельной характеристикой объекта, то можно выделить несколько типов признаков:

- бинарные, принимающие только два возможных значения;

- номинальные, имеющие конечное количество значений;

- количественные, признаки которые могут принимать любые значения.

При этом входящая совокупность признаков называется предикторами, а выходящая — целевыми признаками. Извлекать признаки можно из данных любого типа

К примеру для работы с данными в виде изображений значениями признаков (feature vector) являются сами изображения, таким образом на основе этого можно выделить такие признаки как кластеры пикселей с одинаковыми или близкими значениями, резкие перепады значений пикселей и другие.

В ходе анализа этапа подготовки данных для использования в обучении генеративно-состязательной нейросети (Generative adversarial network) было выявлено, что основной сложностью подготовки выборки (dataset) для генеративно-состязательной нейросети является то, что автоматизация этого процесса не возможна по определению, или же представляет собой задачу для

другого класса нейронных сетей к примеру сверточных сетей (Convolutional Neural Networks).

Даже в этом случае составление основной и конкурирующей выборки возможно только эмпирическим путем на основании визуальных данных. Так как пара генератор-дискриминатор в генеративно-соотязательной сети использует наборы переменных латентного пространства, что предполагает под собой невозможность измерения признаков и выделения их с помощью классических стандартов машинного обучения.

References

1. James, Gareth (2013). An Introduction to Statistical Learning: with Applications in R. Springer. p. 176. ISBN 978-1461471370.

2. Bishop, C.M. (1995), Neural Networks for Pattern Recognition, Oxford: Oxford University Press, p. 372

3. Cohen, S.; Rokach, L.; Maimon, O. (2007). "Decision-tree instance-space decomposition with grouped gain-ratio". Information Sciences. Elsevier. 177 (17): 3592–3612.

4. Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Joshua (2014). «Generative Adversarial Networks».

5. Thaler, SL, US Patent 05659666, Device for the autonomous generation of useful information, 08/19/1997.

6. "A Coevolutionary Approach to Learn Animal Behavior Through Controlled Interaction".: 223–230, Amsterdam, The Netherlands: ACM.

7. Zhuravlev Yu. Y., Riazanov V. V., Senko O. V. Raspoznavanye. Matematycheskye metodы. Prohrammnaia systema. Praktycheskye prymenenyia. – M.: Fazys, 2006.

8. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: Data Mining, Inference, and Prediction. – Springer, 2001. – 533 p. – ISBN 9780387952840.

9. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I Latent Dirichlet allocation (англ.) // Journal of Machine Learning Research : journal / Lafferty, John. — 2003. — January (vol. 3, no. 4–5). – P. pp. 993–1022. – doi:10.1162/jmlr.2003.3.4-5.993.

10. Greene, Jeffrey A.; Brown, Scott C. The Wisdom Development Scale: Further Validity Investigations (англ.) // International Journal of Aging And Human Development : journal. – 2009. – Vol. 68, no. 4. – P. 289–320 (at p. 291). – PMID 19711618.

ANALYSIS OF THE FEATURES OF THE SAMPLE PREPARATION STAGE FOR A GENERATIVE-ADVERSARIAL NEURAL NETWORK

Alexander Zakharov, Selivyorstova Tatjana

Abstract. The authors propose a generalized description of the Data Preparation stage used in Data Mining for training neural networks. The general features of the process of sampling from the general population of data, extraction and generation of features are identified and described. The classification of the sample is given, and the idea is substantiated that the classical approaches to the data preparation stage in data mining are not applicable when preparing data for training and using the generative-adversarial neural network.

Keywords: NEURAL NETWORKS, DATA MINING, DATA PREPARATION, DATASET, GENERATIVE-ADVERSARIAL NEURAL NETWORK, LATENT VARIABLES.