УДК 004.855.5

Проблемы обучения нейронных сетей на несбалансированных данных в задачах распознавания образов.*

А.А. Иванова, Е.Г. Кузнецова, Д.П. Николаев
Институт проблем передачи информации им. А.А. Харкевича РАН

В работе описаны распространенные проблемы построения нейросетевых классификаторов на несбалансированных данных, полученных с сенсоров в режиме реального ограниченного времени. Предложен алгоритм синтеза данных с использованием известных методов обработки изображений для увеличения объема и устранения несбалансированности обучающей выборки. Приведены результаты вычислительных экспериментов, демонстрирующие повышение качества работы классификатора при использовании алгоритма синтеза данных на примере задачи классификации образов символов полей имя-фамилия, полученных с фотографий паспортов РФ. Рассмотрен вопрос построения векторов входных признаков классификатора на основе изображений обучающей выборки, предложен метод нормализации яркости изображений при формировании векторов признаков. Приведены вычислительные эксперименты, показывающие целесообразность использования регуляризации для улучшения обобщающей способности классификатора. Исследован вопрос выбора архитектуры классификатора, обеспечивающей наилучшее качество классификации при существующих ограничениях на быстродействие работы алгоритма в реальном времени. Была экспериментально продемонстрирована целесообразность использования эмпирических методов при решении вопросов выбора архитектуры и параметров обучения нейросетевых классификаторов, показано, что усложнение архитектуры не гарантирует роста качества классификации.

Ключевые слова: машинное обучение, обучение на несбалансированных данных, синтез данных, нейронные сети, регуляризация, обработка изображений, компьютерное зрение, распознавание образов, распознавание символов, контрастирование изображений, бесшовная склейка изображений, сегментация документов.

---

* Работа частично финансово поддержана грантами РФФИ № 13-07-12178, № 13-07-12172

Литература

1. Gary M. Weiss , Forest Provost. The Effect of Class Distribution on Classifier Learning: An Empirical Study. Journal of Artifical Intelligence Research, 2003.
2. A. Estabrooks, T. Jo, N. Jakpowicz. A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence, V. 20, Is. 1, 2004.
3. MathWorks Matlab Documentation official page: http://www.mathworks.com/help/matlab/
4. Haibo He, Edwardo A. Garcia. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, pp. 1263-1284, September, 2009.
5. X.Y. Liu, J. Wu, and Z.H. Zhou. Exploratory Under Sampling for Class Imbalance Systems, Man, and Cybernetics, Part B: Cybernetics. IEEE Trans V. 39, Is. 2, 2008.
6. T. Jo and N. Japkowicz. Class Imbalances versus Small Disjuncts. ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasers, V. 6, Is. 1, pp. 40-49, New York, USA, June 2004.
7. B.X. Wang and N. Japkowicz. Imbalanced Data Set Learning with Synthetic Samples. Proc. IRIS Machine Learning Workshop, Ottawa, 2004.
8. G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasers, V. 6, Issue 1, pp. 20-29, June 2004.
9. P. Domingos. MetaCost: A General Method for Making Classifiers Cost-Sensitive. 5 ACM SIGKDD International Conference on Knowledge discovery and data mining, p. 155-164, New York, 1999.
10. X.Y. Liu and Z.H. Zhou. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. IEEE Trans. Knowledge and Data Eng., 2006.
11. Dan Claudiu Ciresan, Ueli Meier, Luca Maria Gambardella, Jurgen Schmidhuber. Deep Big Simple Neural Nets Excel on Hand-written Digit Recognition. Neural Computation, V. 22, №12, 2010.
12. Patrice Y. Simard, Dave Steinkraus, John C. Platt. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. ICDAR '03 Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003.
13. Larry Yaeger, Richard Lyon, Brandyn Webb. Effective Training of a Neural Network Character Classifier for Word Recognition. Advances in Neural Information Processing Systems, V. 9, Cambridge, 1997.
14. Жуковский А.Е., Тарасова Н.А., Усилин С.А., Николаев Д.П. Синтез обучающей выборки на основе реальных данных в задачах распознавания изображений. Информационные технологии и системы (ИТиС'12): сборник трудов конференции. М., 2012. С. 377-382.
15. P. Perez, M. Gangnet, A. Blanke. Poisson Image Editing. 03 ACM SIGGRAPH 2003 Papers, p.p. 313-318, 2003.
16. J.Sola, J. Sevilla. Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Trans. Nuclear Science, V. 44, № 3, p.p. 1464– 1468, 1997.
17. M. Martins, etc. A new method for multi-texture segmentation using neural networks. Neural Networks. Processin of the 2002 International Join Conference, V. 3, 2002.
18. S. Mato, Y. Yadav. Effectual Approach for Facial Expression Recognition System. International Journal of Advanced Research in Compiter and Communication Engineering, V. 4, May 2015.
19. Krogh A., Hertz J.A. A simple weight decay can improve generalization. Advances in Neural Informatio Processing Systems 4, pp. 950-957, 1992.