|  |  |
| --- | --- |
| Title | Snorkel: Rapid Training Data Creation with Weak Supervision |
| Author | Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Re´ (Stanford University) |
| URL | https://arxiv.org/pdf/1711.10160.pdf |
| Motivation | We present Snorkel, a ﬁrst-of-its-kind system that enables users to train state-of-the-art models without hand labeling any training data. |
| Challenges | 1.we need a way to estimate the unknown source accuracies to resolve disagreements. 2.we need to pass on this critical lineage information to the end model being trained. |
| Techniques | Snorkel has 3 principles:   1. Bring all sources to Bear: The system should enable users to opportunistically use labels from all available weak supervision sources. 2. Training data as the interface to ML: The system should model label sources to produce a single, probabilistic label for each data point and train any of a wide range of classiﬁers to generalize beyond those sources. 3. Supervision as Interactive Programming: The system should provide rapid results in response to user supervision. We envision weak supervision as the REPL-like interface for machine learning.   The architecture of snorkel:   1. writing labeling function 2. Modeling accuracies and correlations 3. Training a Discriminative Model |
| Contributions | 1. flexible interface for Sources: we built an interface layer around the abstract concept of a labeling function (LF). We developed a ﬂexible language for expressing weak supervision strategies and supporting data structures. In a user study, SMEs build models 2.8× faster and increase predictive performance an average 45.5% versus seven hours of hand labeling. 2. Tradeoffs in Modeling of Sources: This paradigm gives rise to previously unexplored tradeoﬀ spaces between predictive performance and speed. This optimizer correctly predicts the advantage of generative modeling over majority vote to within 2.16 accuracy points on average on our evaluation tasks, and accelerates pipeline executions by up to 1.8×. It also enables us to gain 60%–70% of the beneﬁt of correlation learning while saving up to 61% of training time (34 minutes per execution). 3. First end-to-end system for data programming |
| Experiment | Two real world deployments and four tasks on open-source data sets representative of other deployments.  1.Snorkel outperforms distant supervision baselines by an average of 132%.  2.Snorkel approaches hand supervision, coming within 2.11% of the F1 score of hand supervision on relation extraction tasks and an average 5.08% accuracy or AUC on cross-modal tasks, for an average 3.60% across all tasks.  3.Snorkel enables a new and more efficient interaction paradigm: Some biomedical researchers from across the U.S. learned to use Snorkel as part of a two day workshop, and matched or outperformed models trained on hand-labeled training data, showing the eﬃciency of Snorkel’s process even for ﬁrst-time users. |
| Related work | **1.**How to combine different sources: multi-instance learning to reduce the noise in weak supervision sources  (R. Hoffman, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In Meeting of the  Association for Computational Linguistics (ACL), 2011.)  (S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), 2010.)  2.How to estimate the accuracy of labeled sources without a gold standard to compare:  (A. P. David and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society C, 28(1):20–28, 1979.)  3.Crowdsourcing in which workers have unknown accuracy  (N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi. Aggregating crowdsourced binary ratings. In International World Wide Web Conference (WWW), 2013. )  (M. Joglekar, H. Garcia-Molina, and A. Parameswaran. Comprehensive and reliable crowd assessment algorithms. In International Conference on Data Engineering (ICDE), 2015. )  (Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. Journal of Machine Learning Research, 17:1–44, 2016. )  4.Use generative models on with hand-specified dependency structure to label data for specific modalities  (E. Alfonseca, K. Filippova, J.-Y. Delort, and G. Garrido. Pattern learning for relation extraction with a hierarchical topic model. In Meeting of the Association for Computational Linguistics (ACL), 2012. )  (B. Roth and D. Klakow. Combining generative and discriminative model scores for distant supervision. In Conference on Empirical Methods on Natural Language Processing (EMNLP), 2013. )  (S. Takamatsu, I. Sato, and H. Nakagawa. Reducing wrong labels in distant supervision for relation extraction. In Meeting of the Association for Computational Linguistics (ACL), 2012. )  5.Spectral methods  (F. Parisi, F. Strino, B. Nadler, and Y. Kluger. Ranking and combining multiple predictors without labeled data. Proceedings of the National Academy of Sciences of the USA, 111(4):1253–1258, 2014. )  6.Semi-supervised  (O. Chapelle, B. Sch¨olkopf, and A. Zien, editors. Semi-Supervised Learning. Adaptive Computation and Machine Learning. MIT Press, 2009. )  7.Active learning  (B. Settles. Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012. )  8.Transfer learning  (S. J. Pan and Q. Yang. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10):1345–1359, 2010.)  9.Self training  (A. K. Agrawala. Learning with a probabilistic teacher. IEEE Transactions on Information Theory, 16:373–379, 1970.)  (H. J. Scudder. Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory, 11:363–371, 1965. )  10.Co trining  (A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In Workshop on Computational Learning Theory (COLT), 1998. )  11.Data fusion  (X. L. Dong and D. Srivastava. Big Data Integration. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2015. )  (T. Rekatsinas, M. Joglekar, H. Garcia-Molina, A. Parameswaran, and C. R´e. SLiMFast: Guaranteed results for data fusion and source reliability. In ACM  SIGMOD International Conference on Management of Data (SIGMOD), 2017.)  12.Truth discovery  (Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. SIGKDD Explorer. Newsel., 17(2), 2015. )  13.The latent truth model  (B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A Bayesian approach to discovering truth from conflicting sources for data integration. PVLDB, 5(6):550–561, 2012.)  14.how to model user-specified correlations among data sources  (R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, and D. Srivastava. Fusing data with correlations. In ACM SIGMOD International Conference on Management of  Data (SIGMOD), 2014. ) |
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|