

**Supervised Topic Modeling using Hierarchical Dirichlet Process-based Inverse Regression: Experiments on E-Commerce Applications**

**Abstract:**

The proliferation of e-commerce calls for mining consumer preferences and opinions from user-generated text. To this end, topic models have been widely adopted to discover the underlying semantic themes (i.e., topics). Supervised topic models have emerged to leverage discovered topics for predicting the response of interest (e.g., product quality and sales). However, supervised topic modeling remains a challenging problem because of the need to prespecify the number of topics, the lack of predictive information in topics, and limited scalability. In this paper, we propose a novel supervised topic model, Hierarchical Dirichlet Process-based Inverse Regression (HDP-IR). HDP-IR characterizes the corpus with a flexible number of topics, which prove to retain as much predictive information as the original corpus. Moreover, we develop an efficient inference algorithm capable of examining large-scale corpora (millions of documents or more). Three experiments were conducted to evaluate the predictive performance over major e-commerce benchmark testbeds of online reviews. Overall, HDP-IR outperformed existing state-of-the-art supervised topic models. Particularly, retaining sufficient predictive information improved predictive R-squared by over 17.6 percent; having topic structure flexibility contributed to predictive R-squared by at least 4.1 percent. HDP-IR provides an important step for future study on user-generated texts from a topic perspective.

**Existing System:**

First, most supervised topic models require prespecifying the number of topics a priori. Such specification may result in model misspecification when the specified number of topics misrepresent the true underlying topic structure.

Existing supervised topic models treat the proportion of topic mixtures as a reduced dimension representation of the original document and make predictions based on such representations. It is unclear whether these representations contain sufficient predictive information about the response. Statistically speaking, sufficiency entails that the reduced dimension

representation preserves all the information from original documents for making predictions. The missing information in the supervised topic modeling process may diminish the prediction accuracy.

Large text corpora often span several million documents, leaving many supervised topic models unscalable. Most supervised topic models adopt sampling-based inference algorithms, which require hundreds of iterations over each variable across all documents before convergence. Therefore, the scalability of these models is limited.

**Proposed System:**

We propose a novel supervised topic model called Hierarchical Dirichlet Process-based Inverse Regression (HDP-IR). Specifically, the Hierarchical Dirichlet Process (HDP) is a nonparametric topic modeling technique that allows for a flexible number of topics. Inverse Regression (IR) is a sufficient dimension reduction (SDR) technique that makes predictions with provably sufficient information. HDP-IR characterizes the corpus with a flexible number of topics, which prove to retain statistically sufficient information for improved predictive performance.

Moreover, we develop an efficient inference algorithm for model estimation that is capable of examining large-scale corpora with millions of documents. Evaluation of HDP-IR in comparison with the state-of-the-art baseline techniques reveals that both increasing the topic structure flexibility and using sufficient dimension reduction could improve the predictive performance on user-generated review text in e-commerce applications, and the proposed inference algorithm is highly effective in terms of its scalability.