Proposal: Data Mining the Newsvendor Problem

Shawn O'Neil
soneil@cse.nd.edu

ABSTRACT

We propose applying data mining techniques to aid in the prediction of future demands for short lifecycle products, commonly known as the newsvendor problem. In particular, we intend to apply text classification techniques to a web based analogue of the problem's namesake: rather than predicting the sales demand for newspapers, we will attempt to predict the number of comments a story summary will engender on the social news aggregation site `slashdot.org`.

1. INTRODUCTION

The newsvendor problem is that of predicting what the demand will be for some product, so that the appropriate amount can be ordered/manufactured beforehand. This problem is interesting when applied to products where 1) demand is highly uncertain, and 2) the amount to be manufactured needs to be decided well in advance of a short selling season. Many products (often called *short lifecycle* products) have these properties, such as electronics, fashion items, some vaccines, and of course newspapers [1, 4, 2]. If this decision process is to be repeated multiple times, we call it a *multi-period* newsvendor problem.

Traditionally, solutions to the problem make a stochastic assumption about the demands in some way. For example, it is often assumed that the demands will be drawn independently and identically distributed from some known probability distribution. In this case, an order quantity is decided on which maximizes expected profits [3, 5]. Unfortunately, these approaches are often too optimistic in their assumptions that the estimated distributions will be correct.

More recently, worst case approaches have been used to give guarantees on profit (or other measures of success, such as *regret*: the difference between the optimal profit and the actual profit made) under less strict assumptions. For example, one can assume only a lower and upper bound on the demand range and devise a solution which minimizes the maximum regret [6]. These solutions are often too pessimistic to be useful in practice, however.

For this project, we will attempt to predict the demand for an example "product" by analyzing the properties of the product itself. While newspapers are a prototypical short lifecycle item, data about daily newspaper sales is difficult to obtain and associate with the contents of the paper itself.

Instead, we'll focus on an online equivalent. Summaries of news articles which appear on the front page of the news aggregation site `slashdot.org` will serve as our papers, and the number of comments a story engenders will serve as the demand for the that story.

2. DATA SOURCE

Data for this project has already been collected. Story summaries (including title, the editor who posted the story, date and time posted) as well as the number of comments for each have been collected for every front page entry from 2002 to 2007. During data collection, titles and editors' names were considered as regular text words along with the summary. Non-alphanumeric characters were dropped and all words were lower cased.

The intention is to use the 2007 data as holdout for evaluation purposes. Amongst the 2002 to 2006 data, this leaves 34,614 instances to train on.

3. IMPLEMENTATION

For this study of data mining for the newsvendor problem, we'll look at the effectiveness of the Naive Bayes classifier, which is often used for text classification. More importantly, the technique is general enough that it should be applicable to other types of short lifecycle products, even if they aren't text based. On the other hand, the usage of Naive Bayes will necessitate discretizing the output space—the number of comments a story receives.

Because this is text classification in the "bag of words" model, we'll calculate information gain for reach term using the generalized formula given in [7], as this paper makes a compelling argument for information gain as a dimensionality reduction technique in text classification problems.

4. EVALUATION

Whereas in text classification we're usually interested only in classification error rate, for this problem we're mostly interested in the amount of money made using the predictions on the test data. As such, we'll be comparing the results against those given by some stochastic and worst case solutions in terms of monetary regret for the 2007 data.

If the results are positive and Naive Bayes outperforms the stochastic and worst case solutions, this may be considered an advancement on what is usually considered a very difficult problem.

5. MILESTONES

As mentioned, at this point the data has been collected and some interesting summary statistics have been found. By the first milestone, we expect to have the information gain ranking of the terms completed. Further, we hope to have a framework in place for selecting the top N ranking terms, training a model on them, and using the trained model to predict demands for the 2007 data.

By the second milestone, then, we'll be able to reevaluate the model for different values of N and compare the regret suffered to that of the traditional approaches.

6. DELIVERABLES

The final report will include summary statistics of the data set, which will be used as input for the stochastic and worst case solutions. The actual method of class discretization will be discussed, as will the method of turning a course grained class prediction to a demand prediction. Finally, and most importantly, we'll compare the regret suffered by the data mining approach with that of the other solutions.

7. REFERENCES

[1] E. Barnes, J. Dai, S. Deng, D. Down, M. Goh, H. C. Lau, and M. Sharafali. Electronics manufacturing service industry. The Logistics Institute–Asia Pacific, Georgia Tech and The National University of Singapore, Singapore, 2000.

[2] S. Chick, H. Mamani, and D. Simchi-Levi. Supply chain coordination and the influenza vaccination. In *Manufacturing and Service Operations Management*. Institute for Operations Research and the Management Sciences, 2006.

[3] E. L. Porteus. *Foundations of Stochastic Inventory Theory*. Stanford University Press, Stanford, CA, 2002.

[4] A. Raman and M. Fisher. Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research*, 44(4):87–99, January 1996.

[5] H. E. Scarf. A min-max solution of an inventory problem. In *Stanford University Press*, 1958.

[6] C. L. Vairaktarakis. Robust multi-item newsboy models with a budget constraint. *International Journal of Production Economics*, pages 213–226, 2000.

[7] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.