Analysis of Large-Scale Marketing and Social Media Data by Using Machine Learning Algorithm

氏　名（本籍地）	李（り）　銀星（ぎんせい）
学 位 の 種 類	博　士（経営学）
学 位 記 番 号	経博（経営）第 131 号
学位授与年月日	平成 31 年 3 月 27 日
学位授与の要件	学位規則第 4 条第 1 項該当
研 究 科 、専 攻	東北大学大学院経済学研究科（博士課程後期 3 年の課程） 経済経営学専攻
学 位 論 文 題 目	Analysis of Large-Scale Marketing and Social Media Data by Using Machine Learning Algorithm （機械学習を用いた大規模マーケティングとソーシャルメディアデータの分析）
博士論文審査委員	（主査） 教　授　照 井　伸 彦　　准教授　石　垣　　司 教　授　松 田　安 昌

論 文 内 容 の 要 旨

In recent years, big data analysis becomes one of the hot topics in many fields in include marketing field. Particularly, big amount of text data in social media become a popular target for researchers these years. However, most of the researches focus on just use text data to improve precision of forecasting, like stock market, there are still lack of researches about how to combine traditional econometrics model and such kind of unstructured data. On the other hand, text analysis algorithm has also developed a lot in recent years, models like LDA, or Naïve Bayes can extract features from text easily. However, in machine learning field, they are focus on the model precision rather than interpretability, thus, there are few researches about interpretability of text data or result of machine learning algorithms.

The expansion of the Internet has led to massive information posted by consumers online

through social media such as forums, blogs, and product reviews. This provides an opportunity for firms to know consumers' product expectations and evaluations without the need for a direct survey. Using text mining, Grimes (2008) found that 80% of business-relevant information originates primarily as unstructured text.
The purpose of this research is to apply machine learning algorithm, especially text analysis algorithm, to marketing models. I not only focus on the precision of model, but also analysis and interpret the role of text data in the model.

In Chapter 1, we use big text data from twitter, and forecast the sales of a product and make a new economic indicator by using twitter data. We use both interpretable model like logistic regression as well as machine learning model like Support Vector Regression and Neural Network to compare the advantage and disadvantage of each model. The purpose of this research is to build a forecasting model of sales by using Twitter data, which can be easily collected on the internet. Though it is common that marketing variables such as price or promotion information's are used to improve precision of fore-casting, the feature of this research is that we do not involve such kind of marketing variables in our model. For the government, it is very important to make economic indicators to infer economic climate, and forecasting sales for each company or industry by using open source data is one of the tasks. Furthermore, it is possible to detect the change of economic climate which can't be caught by traditional government statistics or industry statistics. Open-source data is also important for the companies which have no price or marketing information or want to grasp the industry situation. In short, this research is not for the companies which can freely use marketing variables of their own, but for those which have no condition to use marketing variables, or aim to build sales forecasting model in objective viewpoints.
In Chapter 2, the active use of daily aggregated store data—POS data that are accumulated automatically at customer check out points is important for most merchandisers, even for those without a membership system. Most traditional methods of analyzing POS data specify market response function after the range of products is limited to a specific category and the number of products is smaller than the number of records (days). This category-based approach is useful when applied to products from well-recognized categories depending on the validity of the assumed categories. However, it cannot be applied to all products in a store, particularly to products that are purchased infrequently over observational periods. It is possible to lose useful information from store data when using the category-based approach. Thus, we apply LDA model, a popular text analysis algorithm, to big POS-data from

supermarket. By applying LDA, we have reduced dimension and divide all products to several market basket. we propose a regression model for high-dimensional sparse data from store-level aggregated POS systems. The modeling procedure comprises two sub-models—topic model and hierarchical factor regression model—that are applied sequentially not only for accommodating high dimensionality and sparseness but also for managerial interpretation.

The volume of high-dimensional data is growing with progress in network technology. These high-dimensional data contain many zeros, that is, sparse in data spaces, as is the case with POS data in our study. That is, the daily number of product items purchased is considerably smaller than that of items displayed in a store, and many items are recorded as having zero sales and information about their price and promotional variables are not recorded, thus producing yet another data entry with zero value. POS data are used to find effective relationships between product items and promotional variables to facilitate efficient management. In these cases, statistical models that accommodating simply high-dimensional data do not work.

We proposed a regression model for high-dimensional and sparse data. In the proposed model, two types of dimension-reduction models, namely, topic model and canonical correlation model, are used sequentially. More specifically, the store-level aggregated sales and marketing mix data are analyzed using a model that combines the topic model with the regression model. First, the topic model decomposes sales of product items into several topics by allocating each unit sale (“word” in text analysis) in a day (“document”) into one of topics based on joint purchase information.

Next, conditional to the topic model estimates, the market response function of an item is estimated by using information regarding variables on the items inside a topic. That is, we construct topic-wise market response functions of an item by using explanatory variables not only of said item but also of the other items belonging to the same topic. In addition, the explanatory variables include not only marketing mix variables but also the numbers of sales of product items in the same topic. We can expect unusual findings from these response functions because sets of related items are not considered in advance and they are not obtained by conventional category-based market response functions.

Finally, we forecast the sales of an item by summing up respective predictors constructed for topics. We compared the model performance with that of conventionally predetermined category regression model to show that (i) our model has the advantage of offering managerial implications obtained from topic-wise regressions defined according to their contexts, and (ii) it has better fit than does the category regression model.

We have basically kept the number of topics for topic model and the numbers of factors for hierarchical factor regression model as fixed in the empirical application, mostly due to the restrictions on computational times. By using R under usual PC environment, it took four days for estimating topic model, and two days for the part of hierarchical factor regression model when numerically calculating inverse of large-scale matrix to recover the structure in original space. We leave this model selection problem for future research by developing more efficient computation procedures, although we currently recognize that the estimation procedure of second sub model's part could be eased if we employ parallel computing for topic wise market response functions.

In Chapter 3, We use Bass model, a famous diffusion model as our basic econometrics model, we use not only sales data but also user-generated online content (or online comments) to describe and forecast the diffusion process of a new product, where online WOM data is plugged into the model as covariates for affecting the change of key parameters over time. From the modeling perspective, our model is characterized as a diffusion model with a time-varying parameter. Then we aim to improve the performance of Bass model by combining topic feature extracted from WoM (word-of-mouth), which are collected from BBS. We not only evaluate the precision of our proposed model, but also interpret the role of text data in our model. we proposed time-varying diffusion models to accommodate social media information. These models belong to the class of systematic variation models and provide useful insights on parameter variations, where we enlarge the information set regarding the diffusion process using product-related BBS text data from before and after the launch of a new product. We use this information based on the recognition that communications in BBS reflect changes in consumer expectations before launch as well as change s in product evaluations of not only the product itself but also the marketing activity and its competitive products. In particular, the communications among potential customers waiting to launch innovative IT products used in our study contain a sort of proxy variable for consumers' expectations before launch, changes in perception and evaluation after launch. Also, it can be possible to observe the change and current consumer's interest.

Our proposed models contain additional variables constituted from BBS text data by applying two approaches for analyzing text data, i.e., sentiment analysis and topic analysis. These variables are used as covariates to explain parameter temporal transitions. These analytical techniques are expected to extract subjective emotional variables and evaluation-based objective variables in BBS, respectively. The empirical study showed that these additional variables lead to an improvement in the model fit

and precision of forecasting by filling a gap between smooth transitions of sales generated by a static diffusion model and realized sales, and they provide the roles of constructed variables in text analysis for the change in model parameters.
In Chapter 4, we extend Bass model from Chapter 3 to multi-generations, and we also use dynamic text analysis method rather than static model, we will prove the effectiveness of dynamic text analysis method when we want to detect the change of customer's demand.

Chapter 5 contains main approaches in my researches. In this chapter, we have applied marketing mix variables as well as social media information to Bass model for multi-generation. We also proposed a new Bass model with prior structure, which is able to forecasting sales for product of new generation without any sales data from this generation. we have proposed Bass model with social media effect, as well as prior structure to parameters. The result shows the effectiveness of social media effect with dynamic labeled topic model both in marketing mix and prior structure. Particularly, it becomes possible to forecast sales of new generation for a technology product before it launched to the market by using prior structure, which is impossible in previous studies.
We also find topic features play different roles in prior structure and marketing mix, it implies we may need to concern about the dynamic feature of unstructured data, as many researches deal with unstructured data by using static text analysis method.

We also applied the result of social media to realistic problem, we prove that besides improving precision of forecasting, text data have more information and we can detect more interesting problems by fully using the latent information from text data.

論 文 審 査 結 果 の 要 旨

ユーザー発信によるSNSやTwitterなど消費者コミュニケーションは、消費者の行動や企業と対話および価値共創に欠かせない情報としてビジネス分野の大規模データの大きな部分を占めている。本論文では、これらのユーザー生成によるテキスト情報のマーケティングへの活用について、大規模テキストデータを分析する機械学習の手法を用いて様々な新しい統計計量モデルを提案している。
まず第1章では、高価格製品のノートパソコンに関する速報性をもつ売上予測を目的にして、同カテゴリーのTwitterデータから潜在ディリクレ配分（LDA）モデルにより抽出した複数の社会ムード変数が売上の先行指標となることを実証し、消費者同士のコミュニケーションに続いて購買行動が起こるとする因果モデルをディープラーニングにより構築することで、速報性が

ありかつ精度の高い予測モデルを提案している。第2章では、実務で普及しているものの活用が十分でない集計POSデータを用いて、LDAモデルを集計数値情報に適用することにより同時購買の文脈をマーケットバスケットとして推定後、各バスケット内の商品売上を同じバスケット内の多数の商品情報によって説明する階層因子回帰モデルを提案し、NP問題を解きながら意外な発見を含む説明変数の市場反応係数の推定を可能とするマーケティングモデルを展開した。第3章では、代表的IT製品であるiPhoneの普及について、ソーシャルメディアの書き込み情報を用いて、ナイーブベイズによる主観的特徴量およびLDAによる客観的特徴量を抽出し、Bassモデルのイミテーション係数および潜在市場規模の時間変化を説明する要因としてモデルに組み込むことにより、テキスト情報が与える動的影響を捉えかつ予測精度が格段に向上することを示した。第4章では、これを複数世代の製品普及モデルに拡張し、動的LDAモデルにより抽出されたソーシャルメディア特徴量がモデルの予測精度に大きく貢献することを実証した。第5章では、さらにモデルを精緻化し、世代間の普及モデルにおけるパラメータ推移を構造化して事前情報としてモデルに組み込むことで新世代製品の発売以前に予測が可能となるモデルを提案し、それが高い予測精度をもつこと、また製品不具合という特定のラベル付きトピックを抽出することで製品不具合問題へ口コミが与える影響を評価できることを示した。
最新の高度な機械学習手法とマーケティングモデルを融合させるモデルを展開し、学術的のみならず実務的にも大きく貢献するモデルを複数提案している。以上により、本論文は博士（経営学）の学位を授与するに値する論文であると認定する。