4G1-01

確率的トピックモデルとSVMを用いたプラント建設における設計変更箇所予測

Design Change Prediction using Probabilistic Topic Model and SVM for Plant Construction

今沢　慶[*1]
Kei Imazawa

[*1]（株）日立製作所　横浜研究所　生産技術研究センタ
Hitachi, Ltd., Yokohama Research Laboratory Manufacturing Technology Research Center

This paper reports a development of CAD change item prediction method from design change notice to predict the amount of materials and additional cost for plant construction.

The developed method consists of LDA and SVM. First, the method estimates topic distribution of design change notice using LDA. Second, the method predicts CAD design change item from topic distribution using SVM with topic selection algorithm.

In conclusion, our method shows higher prediction accuracy than only LDA method by our experiments using actual data (only LDA: 69%, with SVM: 71%).

1. はじめに

プラント建設プロジェクトは，期間が数か月から数年に及び，予算規模が数百億円から数千億円という，長期かつ大規模なプロジェクトである．このような大規模プロジェクトで収益性を確保するためには，現状の予算消化状況を把握し，想定外のコスト発生を迅速に把握し，対策を行うためのコスト管理を行うことが重要な課題の一つである．

しかし，プラント建設プロジェクトでは，多くの部門や企業が関与しているため，予算消化状況を把握することが困難なことがある．特に，建設プロジェクトの最上流にあたる設計部門が設計変更を行った場合，下流工程にあたる製造，調達部門などの多くの部門や企業がその設計変更に対応をする必要があり，関係する全部門の必要コストを把握するまでに時間を要する．そのため，設計変更発生から予算消化状況を把握するまでに，数か月を要することがある．

そこで，設計変更発生時に，必要なコストを予測することを本研究の目的とした．

次に，設計変更が発生した際における対応の流れについて説明する．プラント建設の際に用いられる図面数は，数千枚に及ぶことがあるため，複数の部門が協力し，設計を行っている．そのため，ある特定の部門において，設計変更が発生すると，他部門の設計箇所に変更の影響が伝搬し，設計変更が必要になる場合がある．このような場合，設計変更を実施した部門から，他部門に対して，設計変更指示書が発行され，設計変更指示書を受信した部門では，設計変更に対応する．また，設計部門での設計変更が完了すると，設計変更指示書が製造部門に対して発行され，受信した部門では，設計変更に対応するといった流れで設計変更に対応している．

そのため，設計変更発生時に，図面内の変更項目を予測するためには，設計変更を行った部門が発行した設計変更指示書を入力データとして，図面内の変更項目を予測する必要がある．この設計変更指示書には，設計変更指示内容が自然言語で記述されている．そこで，本研究では，自然言語で記述された設計変更指示内容から，必要なコストを予測する方式の開発に取り組んだ．

プラント建設では，処理を行う必要がある加工作業の作業量や材料などの物量に契約単価を乗じてコストを算出することが多い．そのため，物量を予測することができれば，必要なコストを算出することができる．また，どの項目の物量が変更になるかを予測することができれば，過去の実績から，変更量の確率分布を推定することができる．そこで，本報告では，設計変更指示書から，物量の変更項目を予測する方式を対象とすることとする．

連絡先: 今沢　慶, (株）日立製作所　横浜研究所　生産技術研究センタ，神奈川県横浜市戸塚区吉田町 292，050-3135-2052，050-31350-3412，kei.imazawa.vb@hitachi.com

2. 確率的トピックモデルによる物量の変更項目予測とその課題

設計変更指示書から，物量の変更項目を予測するためには，設計変更指示書に記載された内容が物量のどの項目の話題を扱っているかを特定できれば良い．そこで，本研究では，確率的トピックモデルによって，設計変更指示書の話題を推定することを検討した．

2.1 確率的トピックモデルによる物量の変更項目予測アルゴリズム

Fig. 1 に物量の変更項目予測アルゴリズムのモデリング処理のフローを示す．

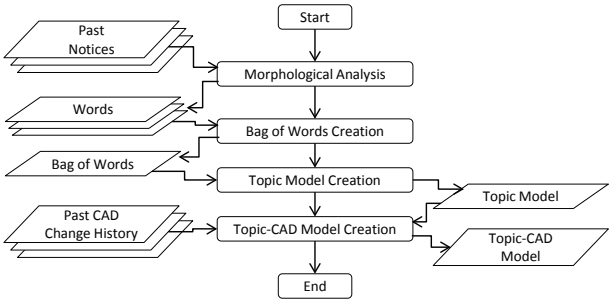

Fig. 1: **CAD Change Prediction Model Modeling Flow**

最初に，過去の設計変更指示書 (Past Notices) に対し，形態素解析処理を行い，単語 (Words) に分割する．

次に，各文書と単語に通し番号を割り付け，各文書ごとに各単語の出現頻度を集計した行列 N(Bag of Words) を生成する Bag of Words 生成処理 (Bag of Words Creation) を行う．Bag of Words のイメージを式 (1) に示す．n_{ij} は文書 i に単語 j が n_{ij} 回出現していることを意味する．

$$N := \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1n} \\ n_{21} & n_{22} & \cdots & n_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ n_{m1} & n_{m2} & \cdots & n_{mn} \end{bmatrix} \quad (1)$$

次に，Bag of Words から，各文書が扱うトピック（話題）を推定するためのトピックモデル (Topic Model) を構築するトピックモデル構築処理 (Topic Model Creation) を行う．本研究では，最も基本的なトピックモデルの一つである LDA(Latent Dirichlet Allocation)[1] を用いた．この LDA は，文章中に登場する単語と文章が取り扱っているトピックとトピックが生起する確率分布 (トピック分布) の 3 つを同時確率分布として推定するモデルで，以下の式 (2) のように同時確率分布を条件付き確率の積に分割できるとみなしたモデルである．

定義 2..1 (LDA : Latent Dirichlet Allocation Model)

$$p(w, z, \theta) = p(\theta|\alpha)p(z|\theta)p(w|z) \quad (2)$$

ただし，w：単語，z：文章が扱っているトピック，θ：トピック分布，α：トピック分布が生起したディリクレ分布のハイパーパラメータとした．

□

本研究では，Griffiths(2004)[2] による方法を用いて，LDA を推定することとした．

最後に，トピックモデルと過去の CAD 変更来歴 (Past CAD Change History) を用いて，各文書が扱うトピックと対応する CAD 変更項目を引き当てるトピック-CAD の項目変換処理を行う．本処理は，以下の仮説に基づいて設計した．

トピックと CAD の項目との関係についての仮説

- トピックと CAD の項目は 1 対 1 に対応する

既開発の方式では，上記の仮説に基づき，過去の CAD の変更来歴 (Past CAD Change History) とトピックモデル (正確には，トピックモデルを用いて算出した各文書のトピック分布) から，トピックと CAD の項目の読み替えテーブル (Topic CAD Model) を構築するトピック-CAD モデル構築処理 (Topic-CAD Model Creation) を行う．

2.2 確率的トピックモデルを適用する上での問題点

前節で述べたアルゴリズムを実データで評価した結果を紹介する．評価データの概要を Table.1 に示す．評価に用いた設計変更指示書数 (Text Sample) は 194 通で，予測対象はパイプルート長さの変更有無 (Yes/No of pipe route length change) である．

また，実験条件を Table.2 に示す．トピック数 (Topic Number) は 20 から 130 の間を 10 刻みで振って実験を行った．本実験では，全データをモデル構築用と精度評価用にランダムに分割して用いた．その分割割合 (Train-Test Ratio) は，1:1 とした．また，モデル構築用と精度評価用のデータをランダムに分割しているため，精度評価結果にばらつきが生じる．そこで，本研究では同じ実験を 10 回繰り返して評価を行った (Repeat Time)．また，10 回繰り返して行った実験結果の中央値を算出して最終的な精度とした (Statistics)．

Table 1: Evaluation Data

Text Sample	194
Target Data	Yes/No of pipe route length change

Table 2: Experiment Condition

Topic Number	20 to 130
Train-Test Ratio	1:1
Repeat Time	10
Statistics	Median

次に，評価に用いた評価指標について述べる．本研究では，2 クラス分類モデルの評価指標として，一般的に用いられる以下の 3 つの指標を導入した．

定義 2..2 (評価指標の定義)

$$\text{TruePositiveRate} := \frac{n_{cc}}{n_{cc} + n_{nc}} \quad (3)$$

$$\text{Precision} := \frac{n_{cc}}{n_{cc} + n_{cn}} \quad (4)$$

$$\text{modFalsePositiveRate} := 1 - \frac{n_{cn}}{n_{cn} + n_{nn}} = \frac{n_{nn}}{n_{cn} + n_{nn}} \quad (5)$$

□

式 (3) で定義する指標は，True Positive Rate(Sensitivity, Recall, Hit Rate) と呼ばれる指標である．また，式 (4) で定義する指標は，Precision と呼ばれる指標である．最後に，式 (5) で定義する指標は，式 (6) で定義する False Positive Rate と呼ばれる指標を微修正したものである．他の 2 つの指標は高いほど性能が良い指標であるのに対し，False Positive Rate(6) は低いほど性能が良いことを意味する指標であるため，理解を容易にするために，式 (5) に示すように，1 から False Positive Rate(6) を引いた値を modified False Positive Rate(5) と定義することにする．これにより，定義 2..2 に示した指標の全ては高いほど性能が良い指標となる．modified False Positive Rate(5) は追加コストを高めに見積もってしまうリスクを評価した指標で，この値が低いと，本来不必要な対策を行ってしまうリスクが高いことを意味する．

$$\text{FalsePositiveRate} := \frac{n_{cn}}{n_{cn} + n_{nn}} \quad (6)$$

Fig.2 に評価結果を示す．横軸がトピック数 (Topic Number)，縦軸が各評価指標の値 (Evaluation Indicator Value) を示している．また，最も性能が良いトピック数を選択した場合の各指標の値を Table3 に示す．

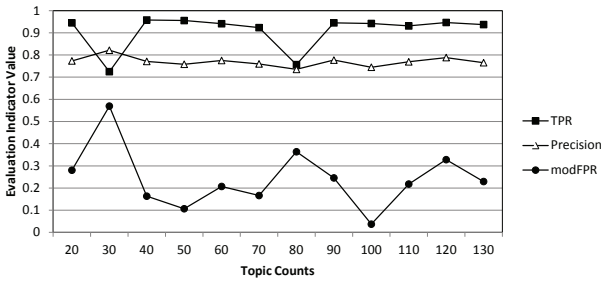

Fig. 2: **Prediction Accuracy of Conventional Method**

Table 3: Best Model Performance

Indicator	Value
True Positive Rate	95%
Precision	80%
modified False Positive Rate	33%

Fig.2 から，本方式では True Positive Rate(TPR) は 80% を超えているが，Precision(Precision) が 80% を下回っている．また，modified False Positive Rate(modFPR) はトピック数が 30 のときを除き，50% に届かない結果となった．また，平均値も 24% に留まっている．このことから，確率的トピックモデルによる予測方式は，ほとんどの設計変更指示書に対して，パイプルート長さが変更になると予測している方式であることがわかる．

以上から，確率的トピックモデルで推定したトピックと CAD の項目は 1 対 1 に対応するという仮説に基づき設計した方式では，充分な予測精度を得られないことが適用上の課題である．

3. 開発方式

本研究では予測モデルを改良するために，トピックと物量の変更項目の関係に関する仮説を以下のように修正した．

トピックと物量の変更項目の関係についての仮説 (改良)

1. トピックモデルによって推定したトピックと物量の変更項目は 1 対 1 対応しない
2. 物量の変更とは無関係なトピックが存在する

これらの仮説に基づき，本研究では物量の変更項目予測方式に次の 2 つの改良を行った．

物量の変更項目予測方式の改良点

改良点 1 トピック分布から物量の変更項目の変更有無を識別モデルによって予測する．

改良点 2 識別モデルの入力データとして用いるトピックを選択する変数選択アルゴリズムを導入する．

上記の改良点についてそれぞれ説明する．

3.1 改良点 1：トピック分布から物量の変更項目の変更有無を予測する識別モデルの導入

トピック分布から特定の物量の変更項目の変更有無を予測するためには，式 (7) の形の関数 f を導入する必要がある．

$$(\text{Yes/No}) = f(k_1, k_2, \cdots, k_K) \tag{7}$$

そこで，本研究では Support Vector Machine (SVM) を導入した．

3.2 改良点 2：トピックを選択する変数選択アルゴリズムの導入

前節で紹介した SVM に対し，Backward 型の変数選択アルゴリズムを導入した．導入した変数選択アルゴリズムの処理手順を Fig.3 に示す．

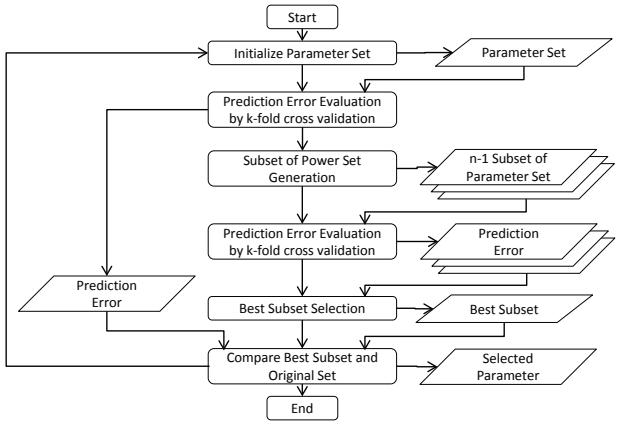

Fig. 3: **Parameter Selection Algorithm Flow**

本方式は Cross Validation により推定した Prediction Error を評価指標とする Greedy 法により，変数選択を行う．

最初に，変数の集合を全集合に初期化する (Initialize Parameter Set)．ここで，全集合とは全ての説明変数からなる集合のことを指す．

次に，全集合で SVM を推定した際の予測誤差を k 分割交差検証法で推定し，$Error(\Theta)$ とする (Prediction Error Evaluation by k-fold cross validation)．

次に，以下の式 (8) で定義する，Θ のべき集合 $\beta(\Theta)$ の部分集合 $\beta_{|\Theta|-1}(\Theta)$ を生成する (Subset of Power Set Generation)．

$$\beta_{|\Theta|-1}(\Theta) := \{\lambda \subset \Theta || \lambda| = |\Theta| - 1\} \tag{8}$$

$\lambda \in \beta_{|\Theta|-1}$ とおくと，λ は Θ から変数を一つ取り除いた変数の集合 ($n-1$ Subset of Parameter Set) になっている．

次に，$\beta_{|\Theta|-1}(\Theta)$ のすべての要素に対して，k-fold cross validation により，以下の (9) で定義する予測誤差の集合 $ErrorSet(\beta_{|\Theta|-1}(\Theta))$ を推定する (Prediction Error Evaluation by k-fold cross validation)．

$$ErrorSet(\beta_{|\Theta|-1}(\Theta)) := \left\{Error(\lambda_i) | \lambda_i \in \beta_{|\Theta|-1}(\Theta), i \in I\right\} \tag{9}$$

但し，$I := \{1, 2, \cdots, |\beta_{|\Theta|-1}(\Theta)|\}$ とおいた．

次に推定した予測誤差が最も小さい $\beta_{|\Theta|-1}(\Theta)$ の要素 λ_0 を特定する (Best Subset Selection)．すなわち，i_0 を以下の式 (10) で定義すると，$\lambda_0 := \lambda_{i_0}$ である．

$$i_0 := argmin_{i \in \{1,2,\cdots,|\beta_{|\Theta|-1}(\Theta)|\}} ErrorSet(\beta_{|\Theta|-1}(\Theta)) \tag{10}$$

この λ_0 が初期化した集合 Θ から変数を一つ取り除いた変数の集合 ($n-1$ Subset of Parameter Set) のうち，最も小さい予測誤差のモデルに含まれる変数の集合 (Best Subset) である．

最後に，初期化した集合 Θ によって構築した SVM の予測誤差と変数の集合 λ_0 によって構築した SVM の予測誤差を比較し (Compare Best Subset and Original Set)，Θ によって構築した SVM の予測誤差の方が小さかった場合，Θ を選択結果 (Selected Parameter) として，処理を終了する．そうでなかった場合，フローの最初に戻り，λ_0 で変数の集合を初期化 (Initialize Parameter Set) し，処理を継続する．

以上の処理により，変更項目予測アルゴリズムの改善を試みた．

4. 開発方式の評価

本章では、2.2 節で用いたものと同じデータ (Table1) を用いた実験により，開発方式の評価を実施した．実験条件も，2.2 節と同様の条件 (Table2) で実施した．実施した実験は，開発方式の精度評価 (実験 1) と 3. 章で立てた仮説を検証するための通常の SVM との精度比較 (実験 2) の 2 つである．

SVM の場合，カーネル関数を選択する必要がある．そこで，本研究では，Lenear，Quadratic，Polynomial，Radial Basis Function(RBF)，Multi Layer Perceptron(MLP) の 5 つのカーネルを用いて評価を行った．

4.1 開発方式の精度評価

評価結果を Fig.4 に示す．横軸がトピック数 (Topic Number)，縦軸が各評価指標の値 (Evaluation Indicator Value) を示している．

Fig.4 (a) は True Positive Rate を示したグラフである．トピック数を増やすと精度が向上する傾向を持ち，トピック数を 120 以上にすると，全てのカーネルで 80% 以上の精度となっている．

Fig.4 (b) は Precision を示したグラフである．トピック数に依存せず，横ばいの推移をしており，多くのケースで 80% 以上の精度となっている．

Fig.4 (c) は False Positive Rate を示したグラフである．トピック数を増やすと精度が減少する傾向を持ち，トピック数が 90 を超えると多くのカーネルで，50% を下回っている．

以上から、True Positive Rate と False Positive Rate がトレードオフの関係にあることが分かり，適切なトピック数を選択する必要があることがわかる．最も性能が良いトピック数とカーネルを選択した場合の各指標の値を Table4 に示す．選択したトピック数は 60，カーネルは Multi Layer Perceptron (MLP) である．True Positive Rate が 80%，Precision が 83%，modified False Positive Rate が 49% であった．

以上から，Precision が確率的トピックモデルによる予測方式が 79% から 83% に比べて改善した．また，modified False Positive Rate も，確率的トピックモデルによる予測方式が 33% であったのに対し，49% に改善した．True Positive Rate に関しては，確率的トピックモデルによる予測方式は，ほとんどの設計変更指示書に対して，物量の変更項目が変更になると予測しているため，提案方式の方が下回る結果となったが，3 指標の平均値は，69% から 71% に改善する結果となった．

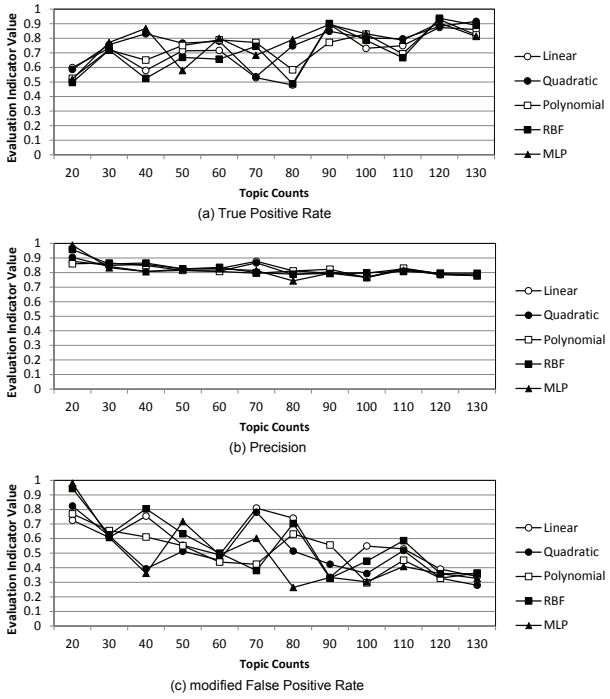

Fig. 4: **Results of Stepwised SVM**

Table 4: Best Model Performance

Indicator	Value
True Positive Rate	80%
Precision	83%
modified False Positive Rate	49%

5. まとめ

本研究では，設計変更発生時に，必要なコストを予測することを本研究の目的に，設計変更指示書から，物量の変更項目を予測する方式の開発に取り組んだ．

確率的トピックモデルのみで物量の変更項目を予測した場合，ほとんどの設計変更指示書に対して，物量の変更項目が変更になると予測してしまう課題があったのに対し，確率的トピックモデルと SVM を組み合わせた方式を検討し，実データで評価した結果，予測精度が 69% から 71% に改善することを確認した．

参考文献

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, vol.3(2003)

[2] Thomas L. Griffiths, Mark Steyvers, Finding Scientific Topics, PNAS, vol.101(2004)