目的変数が範囲で与えられる回帰問題に対するEM法

鹿島久嗣[1]・山崎一孝[1]・西郷浩人[2]・猪口明博[3]

要　　旨

本論文で、我々は、目的変数が範囲として与えられるような回帰問題を考え、この問題への確率的なアプローチを提案する。この問題の最適化問題を直接的に解くことは困難であるが、近似解法としてEMアルゴリズムによる解法を与える。また、提案アプローチの有効性を、価格予測と化合物の活性予測の2つの問題のベンチマークデータセットを用いた数値実験によって示す。

1. はじめに

回帰は、統計、機械学習、およびデータマイニングにおける基本的なタスクの1つであり、実世界においても多くの文脈において現れる重要な問題である。回帰問題では、通常、訓練データとして、入力変数と目的変数のとる値のペアがいくつか与えられると、それをもとに、目的変数の値が未知の入力に対して正しい値を出力を予測するモデルを推定することを目的とする(2章参照)。ここで、訓練データにおいて与えられる目的変数値は、大抵、ある実数値の値をとることを前提としている。しかし、実際の問題においては、目的変数の値についての情報がより曖昧に与えられるような場合があり、このような場合に対応するためには、通常の回帰手法を拡張する必要が生じる。

本論文では、とくに、訓練データの目的変数が「40以上、50以下」などのように、範囲として与えられるような場合の回帰問題を考える。実際に起こりうる問題の例として、以下のようなものが考えられる。

例1：機会損失があるデータからの売り上げ予測

ある商品を、ある店舗に卸して販売したときに、どれだけの数量が売れるかを予測したいとする。商品の説明変数（商品カテゴリや価格など）が入力変数に対応し、その商品が売れる数量が目的変数に対応する。過去の売上データから、商品の説明変数と、売り上げた数量が訓練データとして得られるので、これをもとに、入出力間の関数を推定するのが通常の回帰問題の設定である。しかしながら、過去のデータの中には、本当は卸した数量以上売れるはずだった

[1]日本IBM 東京基礎研究所：〒225-8502　神奈川県大和市下鶴間1623-14
[2]Max Planck Institute for Biological Cybernetics：Spemannstrase 38, 72076 Tubingen, Germany
[2]大阪大学 産業科学研究所：〒567-0047　大阪府茨木市美穂ケ丘8-1

のだが、卸した量が足りなかったため、売り切れてしまった場合が存在する。たとえば、売り上げ上は 2,000 単位となっているが、十分な在庫があれば、2,200 単位は売れたはずであったという場合が存在する。このようなときには、実際売れるはずであった売り上げ数量は「2,000」ではなく「2,000 以上」となるはずであるので、これを「2,000 以上、∞ 以下」の売り上げをもつデータとして、明示的に取り扱うことが望まれる。

例 2：プロジェクトの品質分析

あるプロジェクトに対して、そのプロジェクトの品質を予測するために、プロジェクト品質を表す指標を予測したいとする。プロジェクト品質を表す指標としては、例えば、プロジェクト実行過程におけるトラブルないし、それに準ずる報告の発生件数や、単純に、そのプロジェクトに関わる売り上げなどが考えられる。このとき、プロジェクトの説明変数（プロジェクトマネージャの名前や、参加メンバー数など）が入力変数に対応し、トラブル発生件数が目的変数に対応する。すでに終了したプロジェクトの場合には、そのプロジェクトでのトラブル発生件数は既知であるので、通常の回帰手法においても扱うことができるが、現在もまだ終了していないプロジェクトや、途中で終了してしまったプロジェクトの場合、報告されたトラブル発生件数は、実際に起こる（はずの）件数の下限値となっているはずである。また、担当者が気付かないトラブルや報告義務が無いと勘違いされているトラブルなども存在するため、報告されるトラブル件数は実際のトラブル件数の下限になっている。例えば、「5 件」となっている発生件数は、実際には「5 件以上」となるはずであり、このようなデータは「5 件以上、∞ 以下」の発生件数をもつデータとして、明示的に取り扱う必要がある。

例 3：化合物の活性予測

ある化合物が、薬品としての活性がどの程度あるか予測することは、新規薬剤の効率的な設計において非常に重要な意味がある。この問題を回帰として捉える場合、化合物の特徴を現すいくつかの値あるいは、化合物の分子構造のもつグラフ構造ないし立体構造が入力変数であり、その活性値が出力変数となる。通常の回帰問題では、過去の実験において活性値が計測されている化合物をもとに、活性値の予測モデルを推定する。しかし、実験にかかるコストは無視できないため、実際に活性値が測定されていない化合物も多く存在する。その中でも、エキスパートによって、活性がないと判断されるものがいくつか存在する。この情報を有効に活用するためには、活性がないと判断された化合物については、その活性値が「$-\infty$ 以上、ある小さい活性値（例えば、全体の活性値の大きいほうから 80% の点）」の値をとるものとして、積極的に利用していくということも考えられる。

通常の回帰を行うための手法であれば、最小二乗法による線形モデルの当てはめなどといった基本的なものから、より複雑なベイズ的アプローチまで、各種存在する (Bishop(2006)) が、いずれも本論文で扱うような、目的変数が範囲で与えられるような問題を直接扱えるものはほとんど提案されてない。

そこで、本論文では、3 章において、目的変数が範囲で与えられるような回帰問題に対する確率的なアプローチを提案する。我々は、通常の回帰問題に対する最尤推定や事後確率最大化における目的関数を拡張することによって、範囲をもつ目的変数に対する目的関数を定義する。

この目的関数は、モデルの積分形を含むため、直接の最適化は困難であるが、我々は、近似的な最適化手法として EM アルゴリズムを用いた解法を提案する。

なお、本論文で扱うような回帰問題を陽に扱うことのできる研究としては、Mangasarian et al.(2004) や Le et al.(2006) による、サポートベクトルマシンに制約を入れるアプローチが存在するが、これらは、確率的なアプローチではなく、また、予測時に範囲をもった制約を効果的に取り入れることはできないという点で、我々の研究とは異なっている。

また、最後に、価格の予測と、化合物の活性値予測を扱う 2 種類のデータセットを用いた数値実験を行い、提案アプローチの有効性を検証する。

2. 問題設定

一般に、回帰問題とは、D 次元の実数値ベクトルである入力 $\mathbf{x} \in X = \Re^D$ と、実数値である出力 $y \in Y = \Re$ の間の関係 $f : X \to Y$ を、訓練データと呼ばれる N 組の入出力ペア $E = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$ から推定する問題である。その目的は、出力未知の入力 $\mathbf{x} \notin E$ に対して、正しい出力を予測することである。

本論文であつかう回帰問題は、次の点において、通常の回帰問題と異なる。通常の回帰問題の訓練データが $(\mathbf{x}^{(i)}, y^{(i)})$ のように、出力が 1 つの実数値 $y^{(i)}$ で与えられるのにたいし、本論文では、訓練データが $(\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}])$ のように、出力が「$\ell^{(i)}$ 以上 $r^{(i)}$ 以下」という形で、範囲として与えられる。つまり、訓練データとしては 範囲出力をもつ、N 組の入出力ペア $E = \{(\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}])\}_{i=1}^N$ が与えられる。これは、出力がある特定の値では与えられないが、より曖昧な、ある程度の幅をもって範囲としてなら与えられるような状況を許す、より柔軟な問題設定となっており、1 章で述べたような例をうまく扱うことができる。なお、この設定は、通常の回帰を、$\ell^{(i)} = r^{(i)}$ である特殊な場合として含む点に注意する。これには、出力が点で与えられる訓練データと範囲で与えられる訓練データが混ざっているような場合も含まれる。

さらに、訓練データから学習したモデルを用いて、新しい入力 $\mathbf{x}$ に対する出力の予測を行う際にも、出力の値のとりうる範囲が $[\ell, r]$ の形で与えられており、これを補助情報として用いて予測を行いたいような場合も存在する。

3. 提案アプローチ

3.1 基本的なアイディア

我々は、目的変数の値が範囲で与えられるようなデータにおいて、本当の値は観測されない「隠れ変数」であるとして、EM アルゴリズムによって推定問題を解くアプローチを提案する。EM アルゴリズムでは、大まかには、次のような手続きによって、繰り返しモデルを改善することによって、モデルの推定を行う。

（1） 適当な初期モデルを作成する（例えば、点出力が与えられている訓練データのみを用いて）
（2） 現在のモデルを用いて、範囲出力を持つ訓練データに対して、点出力の「仮置き値」を与える
（3） 点出力をもつ訓練データと、点出力の「仮置き値」をもつ訓練データをあわせて、新しいモデルを推定する
（4） (2) に戻る（モデルが収束するまで繰り返す）

ポイントはステップ 2 であり、EM アルゴリズムにおいては、現在の推定モデルを用いて、隠れ変数の仮置き値を推定するが、これがステップ 2 にあたる。一旦、範囲出力が点出力に置き換えられてしまえば、通常の回帰手法を適用できるため、ステップ 3 は容易に解くことができる。ステップ 4 での繰り返しは、この手法が EM アルゴリズムに基づいているため、必ず、繰り返しによってモデルが悪くなることはなく、また、その繰り返しは収束することが保証される。

3.2 EM アルゴリズムとしての提案手法

前節で述べた手続きを、EM アルゴリズムとして導く。まず、目的変数の値が点で与えられるような場合の最尤推定を自然に拡張することで、最大化するべき目的関数を以下のように定義する。

$$E = \sum_{i=1}^{N} \log \Pr(y^{(i)} \in [\ell^{(i)}, r^{(i)}] | \mathbf{x}^{(i)}) = \sum_{i=1}^{N} \log \int_{\ell^{(i)}}^{r^{(i)}} f(y|\mathbf{x}^{(i)}) dy \tag{3.1}$$

この目的関数を最大化することは、与えられた範囲内に予測が入る確率を最大化しようとしていることになる。通常の、点で出力が与えられるような場合、つまり、$\ell^{(i)} = r^{(i)}$ の場合には、微小な Δ について、$r^{(i)} = \ell^{(i)} + \Delta$ と考えることで、これを含むことができる。

さて、(3.1) を最大化するにあたり、直接の最大化は困難であるため、(3.1) の下界を求め、これを逐次最大化することを考える。$y^{(i)}$ の事後分布の近似 $g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}])$ を導入すると、

$$\begin{aligned} E &= \sum_{i=1}^{N} \log \int_{\ell^{(i)}}^{r^{(i)}} f(y|\mathbf{x}^{(i)}) dy \\ &= \sum_{i=1}^{N} \log \int_{\ell^{(i)}}^{r^{(i)}} f(y|\mathbf{x}^{(i)}) \frac{g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}])}{g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}])} dy \\ &\geq \sum_{i=1}^{N} \int_{\ell^{(i)}}^{r^{(i)}} g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}]) \log \frac{f(y|\mathbf{x}^{(i)})}{g(y|\mathbf{x}, [\ell^{(i)}, r^{(i)}])} dy \end{aligned} \tag{3.2}$$

となる。負等号は、Jensen の不等式を用いた。ここで、f に関係する部分だけ取り出すと、

$$\sum_{i=1}^{N} \int_{\ell^{(i)}}^{r^{(i)}} g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}]) \log f(y|\mathbf{x}^{(i)}) dy \tag{3.3}$$

となり、もしも、モデル f が指数分布族であるとき、すなわち、h と $\boldsymbol{\psi}$ を適当な関数、$\boldsymbol{\eta}$ をパラメータとして、

$$f(y|\mathbf{x}) = h(y) k(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \boldsymbol{\psi}(y)) \tag{3.4}$$

$$1 = k(\boldsymbol{\eta}) \int h(y) \exp(\boldsymbol{\eta}^\top \boldsymbol{\psi}(y)) dy \tag{3.5}$$

のように表される場合には、(3.3) の最大化は、

$$\sum_{i=1}^{N} \int_{\ell^{(i)}}^{r^{(i)}} g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}]) \log f(y|\mathbf{x}^{(i)}) dy \tag{3.6}$$

$$= \sum_{i=1}^{N} \int_{\ell^{(i)}}^{r^{(i)}} g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}]) \log h(y) dy + \sum_{i=1}^{N} \log k(\boldsymbol{\eta}) \exp \left(\boldsymbol{\eta}^\top \int_{\ell^{(i)}}^{r^{(i)}} g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}]) \boldsymbol{\psi}(y) dy \right)$$

となることから、

$$\sum_{i=1}^{N} \log k(\boldsymbol{\eta}) \exp \left(\boldsymbol{\eta}^\top \int_{\ell^{(i)}}^{r^{(i)}} g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}]) \boldsymbol{\psi}(y) dy \right)$$

を最大化すればよいことがわかる。これは、隠れ変数であるところの $\boldsymbol{\psi}(y^{(i)})$ を、現在のモデルにおける条件付期待値

$$\hat{\boldsymbol{\psi}}(y^{(i)}) = \int_{\ell^{(i)}}^{r^{(i)}} g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}])\boldsymbol{\psi}(y)dy \tag{3.7}$$

で「仮置き」して最尤推定を行っていることに相当する。

一旦 f が求まると、条件付確率 g は、

$$g(y|\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}]) = \begin{cases} \frac{f(y|\mathbf{x}^{(i)})}{\int_{\ell^{(i)}}^{r^{(i)}} f(y|\mathbf{x}^{(i)})dy} & y \in [\ell^{(i)}, r^{(i)}] \text{ のとき}, \\ 0 & \text{その他} \end{cases} \tag{3.8}$$

によって求めることができる。

以上より、提案手法における隠れ変数の仮置きと、仮置き値をもとにした最尤推定によるモデル推定は、EM アルゴリズムとして捉えることができる。

3.3 具体的なアルゴリズム

具体的なモデルとして、簡単な線形ガウスモデルを仮定することによって、より具体的なアルゴリズムを導く。モデル f を、

$$f(y|\mathbf{x}, \boldsymbol{\theta}, \sigma) = \mathcal{N}(y|\boldsymbol{\theta}\boldsymbol{\phi}^\top(\mathbf{x}), \sigma\mathrm{I}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\| y - \boldsymbol{\theta}\boldsymbol{\phi}^\top(\mathbf{x}) \|^2}{2\sigma^2}\right) \tag{3.9}$$

のように定義する。 ここで $\mathcal{N}(\cdot|\boldsymbol{\mu}, \Sigma)$ は、平均 $\boldsymbol{\mu}$、共分散行列 Σ の正規分布の確率密度関数を表す。つまり目的変数 y が、平均として線形関数 $\boldsymbol{\theta}\boldsymbol{\phi}(\mathbf{x})^\top$、共分散としては単位行列に定数 σ を掛けた $\sigma\mathrm{I}$ をもつような正規分布に従うものとする。ここで、パラメータは $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_Z)$ および σ である。また、$\boldsymbol{\phi}(\mathbf{x})$ は Z 次元の基底ベクトル（もともと D 次元空間におけるベクトルである $\mathbf{x}$ の、適当な関数 $\boldsymbol{\phi}$ による Z 次元の特徴空間への写像）$\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_Z(\mathbf{x}))$ とする。また、訓練データについて、基底ベクトルを並べた行列 Φ を以下のように定義する。

$$\Phi = \begin{pmatrix} \phi_1(\mathbf{x}^{(1)}), & \phi_2(\mathbf{x}^{(1)}), & \ldots, & \phi_Z(\mathbf{x}^{(1)}) \\ \phi_1(\mathbf{x}^{(2)}), & \phi_2(\mathbf{x}^{(2)}), & \ldots, & \phi_Z(\mathbf{x}^{(2)}) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_1(\mathbf{x}^{(N)}), & \phi_2(\mathbf{x}^{(N)}), & \ldots, & \phi_Z(\mathbf{x}^{(N)}) \end{pmatrix} \tag{3.10}$$

以下、線形ガウスモデルにおける具体的な各ステップを示す。まず、初期化として、点出力を適当に与え、基本回帰アルゴリズムを適用してモデル（厳密にはモデルパラメータ）を得る。あるいは、適当なランダムパラメータを振ってモデルを得る。ここで得られたパラメータを $\hat{\theta}$ とする。そして、以下の E ステップと M ステップを収束するまで繰り返す。

E ステップ

i 番目の訓練データ $(\mathbf{x}^{(i)}, [\ell^{(i)}, r^{(i)}])$ に対して、次の式によって、点出力の仮置き値を求める。

$$y^{(i)} = \frac{\int_{\ell^{(i)}}^{r^{(i)}} y \exp\left(-\frac{1}{2\hat{\sigma}^2}(y - \hat{\boldsymbol{\theta}}\boldsymbol{\phi}^\top(\mathbf{x}^{(i)}))^2\right) dy}{\int_{\ell^{(i)}}^{r^{(i)}} \exp\left(-\frac{1}{2\hat{\sigma}^2}(y - \hat{\boldsymbol{\theta}}\boldsymbol{\phi}^\top(\mathbf{x}^{(i)}))^2\right) dy} \tag{3.11}$$

なお、もしも $\ell^{(i)} = r^{(i)}$ のときには、$y^{(i)} := \ell^{(i)} = r^{(i)}$ となる。

これは、EM アルゴリズムの E ステップで評価するべき期待値 (3.3) において、

$$\int_{\ell^{(i)}}^{r^{(i)}} g(y|\mathbf{x}^{(i)},[\ell^{(i)},r^{(i)}])\log\mathcal{N}(y|\boldsymbol{\theta}\boldsymbol{\phi}(\mathbf{x}^{(i)})^\top,\sigma\mathrm{I})\,dy \tag{3.12}$$

$$= \log\mathcal{N}\left(\int_{\ell^{(i)}}^{r^{(i)}} y\cdot g(y|\mathbf{x}^{(i)},[\ell^{(i)},r^{(i)}])\,dy \,\middle|\, \boldsymbol{\theta}\boldsymbol{\phi}(\mathbf{x}^{(i)})^\top,\sigma\mathrm{I}\right) + const.$$

であることを利用して、隠れ変数であるところの $y^{(i)}$ を、現在のモデルにおける条件付期待値 $E_{y|\mathbf{x},\hat{\boldsymbol{\theta}}}[y|\mathbf{x}^{(i)},[\ell^{(i)},r^{(i)}],\hat{\boldsymbol{\theta}}]$ で置き換えていることに相当する。
(3.11) 式を実際の評価は、閉じた形での解が求まらないため、サンプリングなどを用いて評価を行う。例えば、我々はマルコフ連鎖モンテカルロ法のひとつであるメトロポリス法 (Bishop(2006)) を用いた。

M ステップ

ステップ 2 で得た $y^{(1)},y^{(2)},\ldots,y^{(N)}$ を用いて、通常の回帰アルゴリズムを適用することで、モデルの最尤推定を行い、新しい推定パラメータ $\hat{\boldsymbol{\theta}}$ を得る。

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \sum_{i=1}^{N} \log f(y^{(i)}|\mathbf{x}^{(i)},\boldsymbol{\theta}) \tag{3.13}$$

$\mathbf{y}=(y^{(1)},y^{(2)},\ldots,y^{(N)})$ が与えられると、我々の線形ガウスモデルにおける最尤推定量は、

$$\hat{\boldsymbol{\theta}} = (\lambda\mathrm{I}+\Phi^\top\Phi)^{-1}\Phi^\top\mathbf{y}^\top \tag{3.14}$$

$$\hat{\sigma} = \sqrt{\frac{1}{N}(\hat{\boldsymbol{\theta}}\Phi^\top-\mathbf{y})(\hat{\boldsymbol{\theta}}\Phi^\top-\mathbf{y})^\top} \tag{3.15}$$

によって求まる。なお、ここで λ は正則化パラメータであり、0 以上の定数値とする。

なお、2 章で述べたように、予測時において、新しい入力 $\mathbf{x}$ に対する出力の値のとりうる範囲 $[\ell,r]$ が与えられている場合の予測は、(3.11) 式を使って、$\mathbf{x}^{(i)}:=\mathbf{x}$、$\ell^{(i)}:=\ell$、$r^{(i)}:=r$ として行えばよい。

4. 既存手法および関連研究

既存の手法で、本論文で扱う問題を取り扱おうとした場合に考えられる方法として、以下のような方法および問題点が挙げられる。

アプローチ 1: 目的変数の値が範囲で与えられたようなデータを無視して使わない

アプローチ 2: 範囲のなかで適当な代表点を決めてしまい、目的変数の値として用いる（範囲の平均点や、上記の売り上げ予測の場合では仕入れ量など）

アプローチ 3: 目的変数の値の範囲を制約として用いる (Mangasarian et al.(2004)、Le et al.(2006))。

アプローチ 1 は、範囲として与えられた目的変数の情報を十分には活用できていない点で問題がある。また、アプローチ 2 は、代表点の決め方に任意性がある。本論文での提案手法は、あ

る意味、この代表点を「より正しく」決めているともいえる。また、アプローチ 3 については、有望なアプローチではあるが、確率的な解釈が与えられないこと、また、それゆえ予測時に制約が与えられる場合には、制約の範囲内に入らないものについては、制約を満たすような射影を行うのみで、期待誤差を最小にする予測が行えないという問題がある。[2]

5. 数値実験

ベンチマークデータを用いて、提案手法の性能を検証した。

5.1 価格の予測

まず、UCI Machine Learning Repository に含まれている “housing” データセットを用いて、家の価格を推定する問題で実験を行う。このデータセットは、地域の犯罪発生率や、部屋数など、13 の指標を入力変数として、その家の価格を予測する回帰問題になっており、506 件の事例を含んでいる。

もともと、このデータセットでは、推定すべき目的変数であるところの家の価格は、点出力で与えられているが、以下の手続きによって、今回の問題設定を擬似的に作った。

(1) データの順番を適当にシャッフルする。
(2) 半分（253 件）のデータ（全体の半分）は、点出力のデータとしてそのまま用いる（つまり $\ell^{(i)} = r^{(i)}$ のデータとして）。これを D_{P} とする。
(3) 残り半分のデータは、範囲出力をもったデータとした。 具体的には点出力 $y^{(i)}$ から、以下のようにして範囲出力 $[\ell^{(i)}, r^{(i)}]$ を生成する。ただし、範囲出力の作り方によって、D_{L} と D_{LR} の 2 種類のデータセットを作る。
   (a) 最小値 $\ell^{(i)}$ のみが与えられるデータ D_{L} における範囲出力の生成方法
      i. $0 \sim y^{(i)}/10$ までの一様乱数 $\epsilon_{\mathrm{L}}^{(i)}$ を発生させ、$\ell^{(i)} = y^{(i)} - \epsilon_{\mathrm{L}}^{(i)}$ とした
      ii. $r^{(i)} = \infty$ とする。
   (b) 最小値 $\ell^{(i)}$ と最大値 $r^{(i)}$ の両方が与えられるデータ D_{LR} における範囲出力の生成方法
      i. $0 \sim y^{(i)}/10$ までの一様乱数 $\epsilon_{\mathrm{L}}^{(i)}$ を発生させ、$\ell^{(i)} = y^{(i)} - \epsilon_{\mathrm{L}}^{(i)}$ とする。
      ii. $0 \sim y^{(i)}/10$ までの一様乱数 $\epsilon_{\mathrm{R}}^{(i)}$ を発生させ、$r^{(i)} = y^{(i)} + \epsilon_{\mathrm{R}}^{(i)}$ とする。

データは、90% の訓練データと、10% のテストデータにランダム分割し、これを 30 回繰り返し、30 個のデータセットを生成した。なお、テストデータには、範囲出力を用いていない。

このデータを用いて、線形ガウスモデルにおけるに基づく以下の 4 つの手法を比較した。

(1) 【既存手法 1】点出力のデータ D_{P} のみを使って、通常の回帰を行う
(2) 【既存手法 2】全てのデータを用いるが、点出力のデータ D_{P} 以外のものについては、提案手法において、(3.11) 式を

$$y^{(i)} = \int_{-\infty}^{\infty} y f(y|\mathbf{x}^{(i)}; \hat{\boldsymbol{\theta}}) dy \tag{5.1}$$

で置き換えるとした手法を適用する。この方法は、範囲情報を用いず、単純に予測の期待値で、範囲出力を置き換える方法に該当する。
(3) 【提案手法 L】点出力のデータ D_{P} および、範囲出力のデータ D_{L} を用いて、提案手

[2]なお、本論文においては、このアプローチ 3 との実験比較は行っておらず、これについては今後行っていく予定である。

表 1: “housing”データセットにおける、各手法の平均予測 2 乗誤差の比較。 提案手法のほうが少ない予測誤差であることがわかる。各手法の間の差は、Wilcoxon の符号付順位和検定による p 値で 0.02 以下であった。

手法	既存手法 1	既存手法 2	提案手法 L	提案手法 LR
平均 2 乗誤差	14.18	13.84	12.03	11.37

法を適用する。

（4） 【提案手法 LR】点出力のデータ D_{P} および、範囲出力のデータ D_{LR} を用いて、提案手法を適用する。

なお、(3.10) で、基底には、ガウス基底

$$\phi_i(\mathbf{x}) = \exp(-\alpha \parallel \mathbf{x} - \mathbf{x}^{(i)} \parallel^2) \tag{5.2}$$

を用いた。なお、α はガウス基底のスケールパラメータである。パラメータは、もっとも基本的な既存手法 1 のテスト結果の平均がもっともよかったものを用いた ($\alpha = 0.03$)。

図 1 に、各手法による平均 2 乗誤差を示す。提案手法が最も少ない予測誤差であり、提案手法では、範囲情報を有効に活用できていることが確認できる。各手法の間の差は、Wilcoxon の符号付順位和検定による p 値で 0.02 以下であった。

5.2 化合物の活性予測

次に、化合物の活性値を予測する問題で実験を行う。データは、Saigo et al.(2006) と同じく、National Center for Toxicological Research によって提供されている Endocrine Disruptors Knowledge Base (EDKB) データ[2] から、59 個の化合物が含まれる E-SCREEN アッセイデータを用いた。ここで予測すべき目的変数は、化合物の活性値を表す logRPP 値である。ここでは、化合物の分子構造から、活性値を予測することとし、入力変数としては、化合物をグラフ表現したものに含まれる部分グラフを用いた特徴ベクトルを用いた。具体的には、Inokuchi et al.(2000) の頻出グラフパターン発見アルゴリズムを用いて、全データ (59 個) 中の、5% 以上に含まれる部分グラフを全て数え上げ[3]、これらが化合物のグラフ表現中に出現する (1) か、出現しない (0) かによって、13,600 次元の 2 値ベクトルを生成し、これを入力変数として用いた。

1 章で言及したように、化合物の中には専門家の経験と知識によって、明らかに活性がないと判断されるものがいくつか存在するが、これらの情報をモデル推定において明示的に利用することで予測の改善が期待できる。そこで、この状況を擬似的に作り出すために、まず、全ての化合物を logRPP 値の値によって、大きいほうから並べ、小さいほうから 33% の点 (-0.8421) を見つけた。次に、この値より小さい活性値をもつ化合物のうち、半分をランダムに選び「専門家によって活性がないと判断された化合物」とし、これらの活性値を「$-\infty$ 以上、-0.8421 以下」として用いた。

データは、80% の訓練データと、20% のテストデータにランダムに分割した。これを 30 回繰り返し、30 個のデータセットを生成した。

図 2 に、各手法による平均 2 乗誤差を示す。提案手法が最も少ない予測誤差であり、提案手法では、範囲情報を有効に活用できていることが確認できる。各手法の間の差は、Wilcoxon の符号付順位和検定による p 値で 0.02 以下であった。ここでも提案手法が最も少ない予測誤差であることが確認できる一方、範囲情報を用いない EM 法である既存手法 2 は、既存手法 1 よ

[2] http://edkb.fda.gov/databasedoor.html

[3] 発見されたパターンは 13,600 個、最大サイズ 20 であった。

表 2: “EKBD”データセットにおける、各手法の平均予測 2 乗誤差の比較。 提案手法のほうが少ない予測誤差であることがわかる。各手法の間の差は、Wilcoxon の符号付順位和検定による p 値で 0.02 以下であった。

手法	既存手法 1	既存手法 2	提案手法
平均 2 乗誤差	0.198	0.208	0.190

りもむしろ悪化していることがわかる。

6. おわりに

本論文で、我々は、目的変数が範囲として与えられるような回帰問題を考え、この問題への確率的なアプローチを提案した。提案法における最適化問題の目的関数は、解析的に求めることのできない積分を含むため、これを直接的に解くことは困難であるが、これに対する近似解法として EM アルゴリズムによる解法を与えた。また、具体的なアルゴリズムとして、線形ガウスモデルを用いた場合を示した。また、最後に、家の価格予測および化合物の活性予測の 2 つのデータセットを用いた数値実験によって、提案アプローチの有効性を示した。

今後の発展としては、よりベイズ的なモデル、例えばガウス過程などを同様の問題設定に対して適用できるような拡張を行っていくことが考えられる。また、Mangasarian et al.(2004) などの範囲を制約として用いる手法との比較や、予測時に範囲情報が与えられる場合の予測精度比較などの実験評価も行っていく予定である。

参 考 文 献

Mangasarian, O.L., Shavlik, J.W., and Wild, E.W.(2004). Knowledge-Based Kernel Approximation, *Journal of Machine Learning Research.*, **5**, 1127–1141.

Bishop, C.M.(2006). *Pattern Recognition and Machine Learning*, Springer-Verlag.

McLachlan, G.L., and Krishnan, T. (1997). *The EM Algorithm and Extensions*, Wiley.

Saigo, H., Kadowaki T., and Tsuda, K. (2006). A Linear Programming Approach for Molecular QSAR analysis, *Proc. Mining and Learning with Graphs (MLG 2006)*, Berlin, Germany.

Inokuchi, A., Washio, T., and Motoda, H. (2000). An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, *Proc. The 4th European Conference on Principles of Data Mining and Knowledge Discovery(PKDD)*, 13-27, Lyon, France.

Le, Q.V., Smola, A.J., and Gärtner, T. (2006). Simpler Knowledge-based Support Vector Machines, *Proc. The 23rd International Conference on Machine Learning(ICML)*, Pittsburgh, PA.