1W-05

GIS サンプリングの性能評価

川見　純一†　　三浦 孝夫†

法政大学理工学部創生科学科

東京都小金井市梶野町 3-7-2

1. 前書き

近年、インターネットの普及によりビックデータは増大する一方であり、単なる量の多さだけでなく、意味の変化や状況の変化という質の問題がある。これら大半が文字情報であり、自然言語であるためこれを処理できる方法が求められる。また、自然言語処理の分野ではある事例がどのクラスに属するかを決定する分類問題があり、それらに対する手法は様々あるが、自然言語処理の分野では最大エントロピモデルが散見される。しかし、最大エントロピモデルには素性関数を人手によって設定する必要があり、それらは経験や個人差などが反映されてしまうことや確率モデルを推定するときのパラメータを推定する際に膨大な計算量を必要とするという欠点がある。本研究ではクラス分類において、この膨大な計算量を必要とするすべての候補クラスについての足し合わせを積分計算とみなし、モンテカルロ法による近似手法を提案する。クラス候補の和をサンプリングしたその代表点から近似することにより、学習速度の効率化を行い、学習速度、分類精度をベースラインと比較する。

2. 最大エントロピモデル

2.1 モデル生成

最大エントロピモデルは、学習データから確率分布$P(y|x)$を推定する。モデルパラメータを推定することでモデル推定とし、最大エントロピモデルの条件付確率$P(y|x)$は次式で表せる。

$$P(y|x) = \frac{1}{Z(x,\omega)} \exp\{\phi(x,y)\} \quad (1)$$

$$Z(x,\omega) = \sum_{y \in Y(x)} exp\{\phi(x,y)\} \quad (2)$$

$$\phi(x,y) = \sum_j \omega_i f_i(x,y) \quad (3)$$

制約条件は素性を用いた線形式で表現される。これを表すモデルは式　を満たすモデルの集合Pを定義したとき、式(6)のような確率分布を式(8)の制約条件下で見つける最適化問題をLagrange 法を用いることで導出できる。

$$P = \{P | E_{\hat{P}}[f_j] = E_P[f_j], j \in \{1,2,\dots,n\}\} \quad (4)$$

$$p = arg \max_{p \in P} H(p)\} \quad (5)$$

$$H(P) = -\sum_{x,y} P(x,y) logP(x,y) \quad (6)$$

この結果、解は式(1)、(2)、(3)となる。また素性という文書の特徴を表す値の経験期待値とモデル期待値を一致させるようにパラメータを推定することによって、エントロピと尤度が最大となる。しかし素性関数の設定が分類に大きく影響し、各素性の重みを考慮することで評価を行う。制約が少なく、未知の状況においてエントロピが最大になるような確率モデルを生成するため一様な分布となる。一方で、素性関数の設定を人手で行われなければならない点やGeneralized Iterative Scaling(GIS)やImproved Iterative Scaling(IIS)などの勾配法がベースの計算手法によってパラメータが計算されるため、パラメータ推定計算量が多く、学習速度が遅いという欠点がある。

2.2 GIS

GISやIISは勾配法をベースとした反復スケーリングアルゴリズムで、最大エントロピモデルのパラメータ推定に提案されている。最大エントロピモデルのパラメータは、素性の役割の重要度を表しているため、クラス判別に有効である。モデルの学習では、素性の働きを十分に表現することができる。

しかし実際の過程では、膨大な計算量が必要となる。$E_{\hat{P}}[f_i]$は経験期待値のため、定数であり一度計算すればよい。$E_P[f_i]$ の計算量は$O(|X||Y|)$であるが、GIS ではこれを各素性f_iについて算出し、一回の反復では$O(|X||Y||F|)$ の計算量が必要となる。本研究ではサンプリング方式によるGIS計算の効率化を提案する。提案手法ではクラスyをサンプリングにより生成するためサンプル数をM として、計算量は$(|X||M||F|)$となる。サンプル数M にのみ依存した計算量を要する。近似的に正規化項を求めることで大幅な効率化を行う。

3. モンテカルロ法

モンテカルロ法 では積分計算を期待値の計算とみなす。D次元の変数x を引数にとる関数$f(x)$の領域D における積分

$$I = \int_D f(x)dx$$

を考える。積分Iはxに従う確率分布$p(x)$を用いると式(7)のように表せる。

$$\int_D f(x)dx = \int_D \frac{f(x)}{p(x)} dx \quad (7)$$

$\frac{f(d)}{p(x)}$を$g(x)$とおくと、

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

$$I = \int_D g(x)p(x)dx$$

と表せ、これは確率分布$p(x)$に従う関数$g(x)$の期待値である。このように、積分は期待値計算とみなすことができる。

4. 提案手法

本研究では、モンテカルロ法によるM個のサンプルの和を近似する。正規化項はすべてのクラス候補のスコアを足し合わせているため、これを積分とみなし、モンテカルロ法を適用する。各y_iが分布$p(y)$ に従うがM 個のサンプル$y = (y_1, y_1, \dots, y_M)$を得られたとき、正規化項を期待値とみなし、以下のように次式で表せる。

$$Z(x) = \sum_{y \in Y} \exp\{\phi(x, y)\}$$

$$= \sum_{y \in Y} \frac{\exp\{\phi(x, y)\}}{p(y)} p(y)$$

$$= \frac{1}{M} \sum_{k=1}^{M} \frac{\exp\{\phi(x, y_k)\}}{p(y_k)}$$

この計算はM 回のスコアの計算で終わり、従来の手法よりもM は十分に小さいため、高速な学習速度が期待できる。候補クラスの集合Y の各要素を $(y_1, y_2, \dots, y_{|Y|})$とすると、各クラス$y_i$の生起確率$p(y_i)$で表せる。学習データから経験的なクラス$y_i$ の確率は

$$p(y_i) = \frac{y_i \text{の出現回数} + 1}{|D| + |Y|}$$

で求まる。ゼロ確率問題を避けるため、MAP 推定手法で正規化する。この経験的な生起確率を用いて、サンプリングする。

5. 実験

提案手法を用いて、最大エントロピモデルの学習が行えることを示す。実験はGIS 計算(ベースライン) と、提案手法によるGIS 計算を用いた文書分類を行う。

5.1 実験準備

UCI KDD Archive におけるReuters-21578 Text Categorization Collectionの記事776件を学習データ、記事334件をテストデータとして使用し文書分類の実験を行う。

この実験ではテキストとそのtopic（トピック）を1 件のデータとして使用する。ベースラインには通常のGIS を用いた最大エントロピモデルを使用し、提案手法は逆関数法によるサンプルを発生させ、正規化項を標本平均から計算する手法である。また両手法とも、同様の素性セット及び学習データテストデータを使用する。

5.2 実験結果

ベースラインサンプル数を10、15、20、25、30、35、40、45、50回生成したときの、文書分類の正解率、適合率、再現率、F値、学習時間、正規化項の計算時間、を表9 と表10に示す。ベースライン、提案手法ともに正規化項の計算時間は、パラメータ更新における経験期待値の計算が対象である。

	ベースライン	10	15	20	25
学習時間	122567	52498	77068	94323	120438
正規化項の時間	98608	41567	62794	79171	100953
正解率	0.958	0.963	0.963	0.963	0.963
適合率	0.535	0.535	0.535	0.535	0.538f
再現率	0.535	0.535	0.535	0.538	0.538
F値	0.535	0.535	0.535	0.538	0.538

	30	35	40	45	50
学習時間	162993	162993	185359	206978	223334
正規化項の時間	118654	138708	161242	180388	194518
正解率	0.963	0.963	0.963	0.963	0.963
適合率	0.538f	0.538f	0.535	0.535	0.535
再現率	0.538	0.538	0.535	0.535	0.535
F値	0.538	0.538	0.535	0.535	0.535

5.3 考察

提案手法において、サンプル回数が25個の際、精度が96.3%を占める。クラス数が25個に対し、それに近いサンプル数を生成している。サンプリングは経験的な生起確率分布により近くなるほど、パラメータが正しく変動し、分類精度もベースラインに近くなると考えられる。つまり、クラス数付近のサンプル数を生成した際、サンプリングは生起確率分布に近く、逆関数法を使ったサンプリングの有用性が高いことが分かる。

6. 結論

本研究では逆関数法によるサンプリングを使用し、最大エントロピモデルにおける正規化項を近似的に計算することで、効率的な学習を提案した。通常では時間のかかる膨大な計算量を効率的に行えることが可能であることを示した。また提案手法は多クラスにおける文書分類において、通常の手法の96.3%の精度を保ち42%の時間で学習できることを示した。

参考文献

[1] Junichi Kawami, Takao Miura. “Improving Maximum Entropy Model by GIS.” EJC 2021

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.