別紙４－２（課程博士（和文））

2021年1月8日

情報・知能工学専攻	学籍番号	第143313号
氏名	菊地 真人	

指導教員	梅村 恭司
	青野 雅樹

論文内容の要旨 （博士）

博士学位論文名	統計量の保守的な推定に関する実証的研究

（要旨　1,200字程度）

情報源から得た事象の観測頻度をもとに統計量を推定することは，データを確率的に処理するときの基本操作である．そしてその推定法は，データを用いた工学的応用での有効性を左右する重大な要因になる．現実のデータには高頻度で生じる事象と低頻度で生じる事象が混在する場合があり，この場合でも不偏推定量がよく用いられている．しかし事象の観測が低頻度の場合，不偏推定量は二つの問題を抱えている．第一に，不偏推定量は推定の不確実性が大きい．第二に，不偏推定量は偽の事象を真と誤る第一種過誤，真の事象を偽と誤る第二種過誤を同じ損害とみなすが，実際は一方が他方よりも大きな損害を持つことが多い．そこで本論文では，統計的過誤による損害が小さくなるよう，頻度に応じて推定量を低めに（保守的に）見積もる枠組みを提案した．また扱う統計量としては，条件付き確率と尤度比の二つを推定の対象とした．前者は関係マイニングや確率的言語モデル，後者は多値分類や統計検定などで広く用いられる統計量である．

第1章では，保守的な推定法を提案する背景および本論文の研究目的をまとめた．具体的にはまず，統計量を推定することの重要性と推定に不偏推定量を用いた場合の問題点を説明した．そして，問題点を軽減する方策として保守的な推定法を紹介し，その根本的な考え方を説明した．最後に本論文を成す研究内容を概説した．

第3章では，条件付き確率の保守的な推定法を提案した．この手法は，確率分布の信頼区間を構築し，その下限値を推定値とする．実験では，条件付き確率を用いて新聞記事コーパスから都道府県・市郡間の包含関係を発見した．結果として，提案手法を用いると高・低頻度の両方を効果的に扱い，多くの関係を発見できることを確認した．なお，提案手法を実現するには，低頻度から信頼区間を構築する必要がある．しかし，信頼区間を構築する既存手法は，低頻度から構築した区間に大きな誤差を含む．そこで第2章において，誤差の少ない信頼区間を独自に構築する手法を提案し，条件付き確率の推定にこの手法を利用した．

第4章では，最適化の枠組みによって正則化を導入し，尤度比を保守的に推定する手法を提案した．そして二つの実験で提案手法の有効性と実用性を示した．第一の実験では，尤度比を用いた文字列予測を行い，提案手法の振る舞いと有効性を明らかにした．第二の実験では，半教師有り学習法に提案手法を取り入れ，わずか10個の科学雑誌名をもとに科学ニュース記事から雑誌名を自動抽出した．結果として，提案手法を用いると多数の雑誌名を抽出することができ，提案手法の実用性が示唆された．

第5章では，第4章で提案した尤度比の保守的な推定法を改良し，データに存在しないゼロ頻度のNグラムにも推定値を付与する手法を提案した．この手法では，Nグラム自体の頻度に加え，それを構成する文字や単語に基づく頻度も利用することで，ゼロ頻度のNグラムに対処する．さらに第4章と同様に正則化を導入し，低頻度に対処すると同時に，より情報のある推定値を算出する．そして，固有表現の左Nグラムを尤度比で予測する実験によって，提案手法の有効性を確認した．

第6章では，本論文の研究内容を総括し，今後の展望を述べた．

別紙４－１（課程博士（英文））

Date of Submission (month day, year) : 1 / 8 / 2021

Department of Computer Science and Engineering	Student ID Number	D143313	Supervisors	Kyoji Umemura
Applicant's name	Masato Kikuchi			Masaki Aono

Abstract (Doctor)

Title of Thesis	Empirical Study for Conservative Estimation of Statistics

Approx. 800 words

Estimating statistics based on the observed frequencies of events is a basic operation to process data stochastically. The way of estimation is a significant factor that influences the effectiveness of statistical applications. Real-world data contain frequent and infrequent events, and even in this case, unbiased estimators are used for estimation. However, the estimators have two problems for infrequent events. First, unbiased estimators have a large estimation uncertainty. Second, unbiased estimators regard type I and type II errors as the same damage, but in reality, one often indicates more damage than the other. Therefore, this thesis presents a "conservative" estimation framework. This framework underestimates statistics depending on frequency to reduce the damage caused by statistical errors. In this thesis, two statistics, that is, conditional probability and likelihood ratio, are estimated.

Chapter 1 describes the background for presenting the conservative estimators and the research objectives. First, the importance of estimating statistics and the problems caused by unbiased estimators are explained. Then, the idea of conservative estimation is introduced as a means to alleviate the problems. Finally, the research contents that make up this thesis are outlined.

Chapter 3 presents a conservative estimation method for conditional probabilities. This method builds a confidence interval for the probability distribution and uses its lower limit as an estimator. In the experiments, the estimator is applied to association rule mining tasks, and the results indicate that it can effectively handle both high and low frequencies and discover many rules. To realize a conservative estimation, it is necessary to construct confidence intervals from low frequencies. However, existing construction methods include large errors in the intervals constructed from low frequencies. Therefore, Chapter 2 presents a new method for constructing a confidence interval with a small error, and the method is used to estimate conditional probabilities.

Chapter 4 presents a conservative estimation method for likelihood ratios (LRs). This method introduces regularization in an optimization framework and achieves conservative estimation. Two experiments demonstrate the effectiveness and practicality of the method. The first experiment is a string prediction task using LRs, and the results clarify the behavior and effectiveness of the conservative estimator. In the second experiment, the presented method is incorporated into a semi-supervised learning method, and scientific journal names are automatically extracted from scientific news articles based on only 10 journal names. As a result, many journal names can be extracted, suggesting the practicality of the conservative estimation.

Chapter 5 presents an LR estimation method to provide informative estimates for low-frequency and zero-frequency (i.e., unobserved) n-grams. This method deals with zero-frequency n-grams by using the frequencies based on the letters and words that compose an n-gram in addition to the original n-gram frequency. Furthermore, this method also introduces regularization to deal with low frequencies. In the experiments, left n-grams of the named entities are predicted using LRs, and the results demonstrate the effectiveness of the proposed estimator.

Chapter 6 provides the overall conclusion and describes the future work.