氏名	張　天豫
学位の種類	博士（応用情報科学）
学位記番号	博情第　49　号
学位授与年月日	平成　30 年　3 月　22 日
学位授与の要件	学位規則第 4 条第 1 項該当（課程博士）
論文題目	Research of Applying Machine Learning Methods to Outlier Detection in Wireless Sensor Networks
論文審査委員	（主査）教授 中本　幸一 （副査）教授 申　吉浩 （副査）准教授 大島　裕明

学位論文の要旨

Wireless sensor networks (WSNs) can be flexibly deployed and used to collect data from various environments. By analyzing the collected data, WSNs can be used for such tasks as environment monitoring, disaster prevention, and event detection. However, collected datasets sometimes contain outliers, which obviously reduce the accuracy of data analysis and the performance of the WSN (e.g., the outliers may trigger a false alarm that generates unnecessary fears). Therefore, removing such outliers before analyzing the collected data is necessary to improve the performance of the WSNs. Outlier detection is the process of data analysis. In WSNs, outlier detection involves two major approaches, which are defined as “centralized” and “distributed.” Our proposed algorithms use the distributed approach, which enables every sensor node to detect outliers on its own and locally. Therefore, in this doctoral thesis, we propose three algorithms for distributed detection of outliers, all based on machine learning. The first and second algorithms are based on supervised and unsupervised learning, respectively. The third is designed to improve the performance of clustering algorithms categorized as unsupervised learning.

The first algorithm is based on supervised learning. It first uses training data to train a classifier on a powerful base node and then distributes this classifier into every remote sensor node. Moreover, this method is founded on a widely used assumption in WSNs in which the entire deploying environment has the same condition. Using this assumption, we can simply gather the training data by defining a normal situation in such an environment. In this simple case, using a user-determined threshold is sufficient. For example, if a WSN is deployed to monitor the temperature of a store, we can determine a threshold based on the previously collected normal data. The threshold can be used to detect those data that represent an outlier. However, when WSN-collected data points contain multiple features, the method based on a threshold is not appropriate. Because a situation involving a data point, such as a normal situation or outlier, is commonly determined by multiple features, when data points have multiple features, a decision bound is used to detect the outliers. In our study, with the help of training data, we used a logistic regression function to calculate the decision bound for multiple-feature outlier detection. In simulations in which the collected dataset contains a different ratio of outliers, this

algorithm can provide a believable decision bound. Moreover, the training of the algorithm is executed on the sink node, whereas outlier detection is executed on the sensor nodes.

Although the support vector machine (SVM)-based method can provide an inspired performance under the aforementioned assumption, this assumption is not reasonable when the deploying environment is very large, as this type of situation is no longer normal. For example, based on their different functions, all rooms in a building have their own sub-environments. Therefore, the normal situation standard of the rooms is different. In this case, preparing training data must involve labeling the situation of considerable data in many sub-environments. Moreover, the sub-environment situation commonly changes over time. For example, people regularly enter or leave a room, which makes the work of preparing training data more difficult. All of these reasons make preparing training data particularly difficult. As a consequence, unsupervised-learning-based methods, which are free of training data, are sensible for solving such problems.

The first unsupervised machine learning algorithm we propose is based on the mean-shift algorithm, which is a clustering algorithm, and we introduce two new distance and anchor data points in our algorithm for outlier detection. In general, clustering algorithms are usually used when data lack additional information or prior experience (e.g., data point labels in the training data). Clustering algorithms are then used to divide a dataset into clusters, where a cluster is defined as a set of data points having similar properties, such as density, in many data analysis tasks. Moreover, we can create a criterion for outlier or event detection by utilizing the results of clustering. In this study, we tested our algorithm on a real dataset from Intel Lab, and it generated an ideal result. Specifically, it found outliers with a low false positive rate and high recall. For generality, we also tested our method on different synthetic datasets.

A clustering algorithm has a drawback in that the number of calculations is high and clustering accuracy sometimes is poor. To enable the clustering algorithm to be faster and more accurate, we propose a new algorithm called the peak searching algorithm (PSA). Traditional clustering algorithms such as EM and k-means algorithms require extensive iterations to form clusters, which result in slow processing speeds. In addition, clustering results are less accurate because of the manner in which clusters are formed. To address these problems, we first propose PSA, which uses Bayesian optimization to find the peaks of the probability of the dataset to enable clustering algorithms to be faster and more accurate, and we then adapt PSA to include the EM and k-means algorithms (PSEM and PSk-means, respectively). Simulation results show that our proposed PSEM and PSk-means algorithms considerably decreased the number of iterations of clustering to 6.3 times (a reduction of 1.99) and improved clustering accuracy to 1.71 times (an increase of 1.69) as compared to the traditional EM and enhanced version of k-means (k-means++) on both synthetic datasets. Moreover, in a simulation of WSN application for outlier detection, PSEM correctly found the outliers in the real dataset. In addition, it decreased iterations by 1.88 times and had a maximum accuracy gain of 1.29 times.

論文審査の結果の要旨

近年，センサーを具備したセンサーノードとこれを無線通信で接続した無線センサーネットワークが注目を浴びている． センサーから集めたデータは必ずしも正常なデータばかりではなく，異常値(outlier)が含まれる．異常値を検出し除去することはセンサーデータの正確性を向上させ，センサーデータを利用したサービスの品質向上になる．しかし，どういう値が異常値なのかは環境の変化などにより変わってくる．例えば，気温では夏と冬では正常な値も異なっているので異常値も変わってくる．また，そうした環境変化は多様であり，フレキシブルで軽量な異常値検知アルゴリズムが必要となっている．本博士論文は機械学習アルゴリズムを利用して，この問題の解決を試みている．

本博士論文では，2 章で異常値検知の手法や機械学習，特に無線センサーネットワークにおける関連研究を述べている．3 章で教師ありアルゴリズムを適用した実験を予備実験として実施し、その限界を述べている．教師ありアルゴリズムの一つである Logistic Regression では，学習データに正常，異常のラベル付けを行う必要があり，環境の変化に適応が難しい点である．4 章で，教師なし機械学習アルゴリズムの一つである Mean-shift アルゴリズムによりクラスタを構成し，異常値検知ができることを示した．Mean-shift アルゴリズム既存の方法に比べて，異常値の割合が 25%と多い場合でも False Positive Rate は 2%程度と低いという結果を生成したデータと実際のセンサーネットワークのデータを利用したシミュレーションにより示した．ただし，既存の教師なしのクラスタリングアルゴリズムには計算時間が長くなる，クラスタ構成の正確さが低減する場合があるという問題がある． 5 章ではこれらの問題を解決するためのピーク探索アルゴリズム(PSA)を述べている．PSA はデータセットの確率のピークを見つけるために Bayesian 最適化を利用している． EM アルゴリズムや k-means アルゴリズムに利用した場合に，オリジナルのアルゴリズムに比べて，これも同様のシミュレーションにより正確さにおいて 2 倍，実行時間において約 1/3 になることを示した．またセンサーネットワークでも利用される小型組込みシステム（RasberyPi）で 1 秒未満で処理できることを示し，実際の環境でも十分利用可能なアルゴリズムであることを示している．

無線センサーネットワークの利用場面は今後も広がると考えられ，異常値検知はその発展を大きく推し進める技術の一つである．本研究のアプローチは極めて実践的であり，社会や産業での進展に貢献することが大であり，本博士論文の実用面での価値も大きいということができる．以上を総合して本審査委員会は，本論文が博士（応用情報科学）の学位授与に値するものと全員一致で判定した．