機械学習手法と画像処理の高度応用に関する研究：交通、天気、パターン認識、ノイズフィルタリング

Summary

In today's digital world, classification and prediction algorithms with the easy access to the required information from huge amount of data from different areas (including medicine, biology, transportation, and weather) plays a critical role in environmental awareness systems.

The second section of this dissertation is the analysis and comparison of classification accuracy between two machine algorithms: Naïve Bayes and *k*-NN was addressed by utilizing the traffic data collected from Ojana junction, Okinawa, Japan. From this analysis, we found that *k*-NN (with 100% accuracy) has a stronger resistance to unbalance datasets than Naïve Bayes (with 98% accuracy). But, both algorithms still suffered from the lazy execution of matching newcomers with each instance throughout the entire dataset. Then, we applied multi-threading approach to Naïve Bayes and *k*-NN. Although, the approach reduced much classification time of both algorithms, unfortunately, the creating and deleting threads also consumes much time. Thus, we mainly focused on these issues in the approach (PRD-*k*NN).

In the third section, this dissertation addressed a hybrid approach for traffic flow estimation by selecting the most suitable approach with the means of integrating proficiencies from the three prediction models: multinomial logistic regression, decision trees, and support vector machine (SVM). The experimental results with the test cases and simulations showed that the hybrid approach with 97% accuracy is more effective than individual methods (decision trees with 90%, SVM with 96%, and multinomial logistic regression with 89%).

In the fourth section, this dissertation presented plurality rule-based density and correlation coefficient-based clustering technique for *k*-NN (PRD-*k*NN) aimed at upgrading the performance of normal *k*-NN by leveraging computational time and improving classification accuracy. By using the real datasets (on breast cancer, breast tissue, heart and the iris) from the UCI machine learning repository, the experimental results of classification accuracy proved that the proposed approach with 99.2% is more efficient than classical *k*-NN with 94.4% and DPC-KNN-PCA with 81.05%, while in the processing time performance, the proposed approach improved the total average 3.53 (353%) times over the classical *k*-NN. The PRD-*k*NN based on the feature selection approach to choose the best

appropriate feature as well as to avoid a very time-consuming job. However, some unrelated or weak attributes still supports making correct decisions for a long-term approach.

In the fifth section, the new approach, named as dual-*k*NN, was presented to avoid some unrelated or weak attributes and to improve the robustness over varying *k* values. By conducting experiments on real datasets from UCI, dual-*k*NN with 71% accuracy is more suitable than classical *k*-NN with 69% accuracy. Besides, the dual-*k*NN with 98% occupies a higher accuracy than CB-*k*NN with 97% for breast cancer, and for heart datasets, the dual-*k*NN with 96% achieves a better accuracy than D-*k*NN with 81%, DPC-*k*NN with 81%, and DPC-*k*NN-PCA with 83%.

In the sixth section, a short-term localized rainfall prediction system based on the rainfall-level data model was presented by applying the dual-*k*NN. The experimentation by using 2011, 2013, and 2014 datasets collected from WITH radar installed on the rooftop of Information Engineering, University of the Ryukyus confirmed that the dual-*k*NN is more efficient than classical *k*-NN. In this research, the experiment is divided into three sections (for the whole rainfall cycle, only for the rainfall level in growth condition and for the decaying condition). To express in details, the dual-*k*NN occupies 95% for the whole cycle, 95% for the growth conditions, 89% for decaying condition, while classical *k*-NN has 93% for the whole and growth condition and 80% accuracy for the decaying condition.

In the seventh section, we proposed Regional Distance-based k-NN (RD-*k*NN), aimed at improving the performance of *k*-NN and the efficiency of RD-*k*NN was shown by comparing with the normal-*k*NN with the three distance measures (Euclidean distance, Cosine distance, and Mahalanobis distance) based on the real data sets from UCI machine learning repository. Euclidean based RD-*k*NN gain the highest classification accuracy (87%), Cosine similarity based RD-*k*NN achieved 84%, and mahalanobis based RD-*k*NN attained 58%.

In the eighth section, this research addressed a kd-tree-based dual-kNN to overcome the issues of pure dual-*k*NN. As the dual-*k*NN was a renew version of normal *k*-NN, the dual-kNN still suffered from a time complexity problem when matching each newcomer to the entire dataset. By conducting experiments on real datasets and comparing this algorithm with two other algorithms (pure dual-kNN and normal k-NN), the kd-tree-based dual-kNN was more effective and robust approach for pattern classification.

In the ninth and tenth section, this dissertation investigated the attribute and class (mislabeled) noise-tolerant levels of dual-kNN by comparing with the noise resistance levels of normal k-NN, PRD-kNN, logistic regression, and neural network. In this case, the algorithms occupied the stronger attribute noise resistance with the average resistance level (88%) than class (mislabeled) noise with 68% resistance. To express in details, dual-kNN achieved the highest resistance level for both attributes noise (91%) and mislabeled noise (69%) respectively.

In the eleventh section, this research introduced a noise reduction algorithm named an adaptive morphological operation for high performance weather image processing. The computation of adaptive approach is $(N^2S^2 + N^2)$ and the conventional approach takes $(2\ N^2\ S^2)$. Because $(2\ N^2\ S^2) < (N^2S^2 + N^2)$, the adaptive approach is more efficient than conventional approach. Besides, for the accuracy, the adaptive approach maintained the important region 1.05 times more than the conventional approach, and similarity, weighted value adaptive approach protected the essential pixels 1.07 times more than weighted value conventional approach.

ABSTRACT

Machine learning is an algorithm for learning concepts and seeking structural patterns in data. With automatically improving experience through those learnings, estimating future events could be done successfully. Thus, machine learning has taken a vital role in the field of medical diagnosis, financial, weather, traffic jam, speech recognition, etc. To build a reliable and accurate prediction/classification model the core tasks of the machine learning field are classification, regression, clustering, density estimation, and dimensionality reduction, etc. Another key thing to remember is that machine learning algorithm performs the prediction/classification work based on very large training examples as well as requires higher data quality to achieve a good approximation. Therefore, from the point of computational issue, the algorithm needs to exploit much time to make a decision for each newcomer by taking greedy search through datasets, and from the point of accuracy issue, if the data is inadequate then the algorithm will fail to make a correct classification too.

Thus, this dissertation aims to boost the robustness of the algorithms and reduce the computational time by introducing advanced machine learning algorithms for pattern classification, rainfall level prediction, traffic jam and noise filtering for rainfall radar image developed in this thesis.

As a first proposed approach for an intelligent transportation system, this dissertation addresses a multi-threaded machine learning system using multisensory fusion and a hybrid approach, which stands on accuracy comparison of three prediction models: multinomial logistic regression, decision trees, and support vector machine (SVM), for traffic flow estimation. Then, this dissertation introduces plurality rule-based density and correlation coefficient-based clustering for *k*-NN (PRD-*k*NN), dual-*k*NN, regional distance-based *k*NN classification (RD-*k*NN), and kd-tree-based dual-*k*NN to improve the performance of *k*-NN algorithm according to the lack of high-speed computation, high robustness, and maintenance of accuracy

for different k values. The results experimentally confirm that the effectiveness of those algorithms by comparing with normal k-NN, density peaks clustering based on k-NN and principal component analysis (DPC-KNN-PCA), logistic regression, and neural network, and conducting experiments on real datasets from UCI machine learning repository, and rainfall radar images from the WITH small-dish aviation radar installed on the rooftop of Information Engineering, University of the Ryukyus.

As we know, machine learning algorithms are powerful tools for a variety of application domains, giving widely divergent dimensions such as reliability, precision, robustness, high-speed solution, etc. Likewise, the other critical dimension that a well-designed learning algorithm should occupy is a high strength of unpredictable and phenomenal noise. For this critical dimension, this research intends to investigate the attribute and class (mislabeled) noise-tolerant level of dual-kNN by injecting different noise levels. The empirical experimentations describe that dual-kNN has a higher attribute and class noise-resistant level than normal k-NN, PRD-kNN, logistic regression, and neural network.

It is never possible to be able to collect a perfect real dataset due to data corruption because of the sensor or acquisition devices, and data transmission. The data corruption constantly forces the learning algorithms to struggle with the prediction or classification works and faces performance degradation in terms of mainly prediction accuracy. Thus, this dissertation presents a noise filtering algorithm as the last proposed approach of this research, named as an adaptive morphological operation for high-performance weather image processing. The experimental results based on 2011, 2013, 2015, and 2016 datasets, confirms that the adaptive approach is more efficient than the conventional morphological operations.