Faculty of Engineering and Information Technology
University of Technology, Sydney

Coupled Similarity Analysis in Supervised Learning

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Chunming Liu

October 2015

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

_____

Acknowledgments

First and foremost, I would like to express my deepest appreciation to my supervisor, Prof. Longbing Cao, for his professional guidance, selfless help and continuous support throughout my PhD study and research. I feel very lucky to have had him as my advisor and I will always remember our many discussions and his invaluable ideas. I sincerely thank him.

I am grateful to my colleagues and friends, Can Wang, Junfu Yin and Xuhui Fan, for their selfless support, and especially to Can Wang, for her advice regarding the technical aspects of my thesis.

My sincere gratitude is extended to my team leader, Zhigang Zheng, for his ongoing support in the Australian Taxation Office project. I am also grateful for the excellent ongoing help that I received from my colleagues and team members, David Wei, Mu Li, and Wei Cao.

I would like to thank all the staff in our Advanced Analytics Institute (AAI). Without their generous support this dissertation would not have been possible.

Last but not the least, I would like to thank my family: my wife and my daughter, for their unconditional love and support throughout my PhD candidature.

Chunming Liu

May 2015 @ UTS

Contents

List of Figures

List of Tables

Abstract

In supervised learning, the distance or similarity measure is widely used in a lot of classification algorithms. When calculating the categorical data similarity, the strategy used by the traditional classifiers often overlooks the inter-relationship between different data attributes and assumes that they are independent of each other. This can be seen, for example, in the overlap similarity and the frequency based similarity. While for the numerical data, the most used Euclidean distance or Minkowski distance is restricted in each single feature and assumes the features in the dataset have no outer connections. That can cause problems in expressing the real similarity or distance between instances and may give incorrect results if the inter-relationship between attributes is ignored. The same problems exist in other supervised learning, such as the classification tasks of class-imbalance or multi-label. In order to solve these research limitations and challenges, this thesis proposes an insightful analysis on coupled similarity in supervised learning to give an expression of similarity that is more closely related to the real nature of the problem.

Firstly, we propose a coupled fuzzy kNN to classify imbalanced categorical data which have strong relationships between objects, attributes and classes in Chapter 3. It incorporates the size membership of a class with attribute weight into a coupled similarity measure, which effectively extracts the inter-coupling and intra-coupling relationships from categorical attributes. As it reveals the true inner-relationship between attributes, the similarity strategy we have used can make the instances of each class more compact when

measured by the distance. That brings substantial benefits when dealing with class imbalance data. The experiment results show that our supposed method has a more stable and higher average performance than the classic algorithms.

We also introduce a coupled similar distance for continuous features, by considering the intra-coupled relationship and inter-coupled relationship between the numerical attributes and their corresponding extensions. As detailed in Chapter 4, we calculate the coupling distance between continuous features based on discrete groups. Substantial experiments have verified that our coupled distance outperforms the original distance, and this is also supported by statistical analysis.

When considering the similarity concept, people may only relate to the categorical data, while for the distance concept, people may only take into account the numerical data. Seldom have methods taken into account the both concepts, especially when considering the coupling relationship between features. In Chapter 5, we propose a new method which integrates our coupling concept for mixed type data. In our method, we first do discretization on numerical attributes to transfer such continuous values into separate groups, so as to adopt the inter-coupling distance as we do on categorical features (coupling similarity), then we combine this new coupled distance to the original distance (Euclidean distance), to overcome the shortcoming of the previous algorithms. The experiment results show some improvement when compared to the basic and some variants of kNN algorithms.

We also extend our coupling concept to multi-label classification tasks. The traditional single-label classifiers are known to be not suitable for multi-label tasks anymore, owing to the overlap concept of the class labels. The most used classifier in multi-label problems, ML-kNN, learns a single classifier for each label independently, so it is actually a binary relevance classifier. As a consequence, this algorithm is often criticized. To overcome this drawback, we introduce a coupled label similarity, which explores the inner relationship between different labels in multi-label classification according

to their natural co-occurrence. This similarity reflects the distance of the different classes. By integrating this similarity with the multi-label kNN algorithm, we improve the performance significantly. Evaluated over three commonly used verification criteria for multi-label classifiers, our proposed coupled multi-label classifier outperforms the ML-kNN, BR-kNN and even IBLR. The result indicates that our supposed coupled label similarity is appropriate for multi-label learning problems and can work more effectively compared to other methods.

All the classifiers analyzed in this thesis are based on our coupling similarity (or distance), and applied to different tasks in supervised learning. The performance of these models is examined by widely used verification criteria, such as ROC, Accuracy Rate, Average Precision and Hamming Loss. This thesis provides insightful knowledge for investors to find the inner relationship between features in supervised learning tasks.