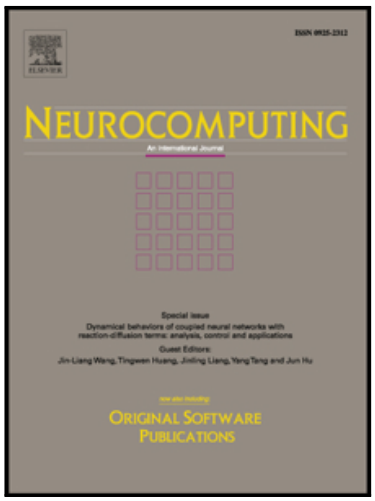

Self-training semi-supervised classification based on density peaks of data

Di Wu , Ming-Sheng Shang , Xin Luo , Ji Xu , Hu-Yong Yan , Wei-Hui Deng , Guo-Yin Wang

PII: S0925-2312(17)30960-8
DOI: 10.1016/j.neucom.2017.05.072
Reference: NEUCOM 18490

To appear in: *Neurocomputing*

Received date: 29 April 2016
Revised date: 24 May 2017
Accepted date: 27 May 2017

Please cite this article as: Di Wu , Ming-Sheng Shang , Xin Luo , Ji Xu , Hu-Yong Yan , Wei-Hui Deng , Guo-Yin Wang , Self-training semi-supervised classification based on density peaks of data, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.05.072

Self-training semi-supervised classification based on density peaks of data

Di Wu[a,b], Ming-Sheng Shang[a], Xin Luo[a], Ji Xu[a,c], Hu-Yong Yan[a], Wei-Hui Deng[a], and Guo-Yin Wang[a,*]

a Chongqing Key Laboratory of Big Data and Intelligent Computing, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, P.R. China
b University of Chinese Academy of Sciences, Beijing 100049, China
c School of Information Science and Technology, Southwest Jiaotong University 610031, Chengdu, China
**Corresponding author: wanggy@ieee.org*

***Abstract*—Having a multitude of unlabeled data and few labeled ones is a common problem in many practical applications. A successful methodology to tackle this problem is Self-Training semi-supervised classification. In this paper, we introduce a method to discover the structure of data space based on find of density peaks. Then, a framework for Self-Training semi-supervised classification, in which the structure of data space is integrated into the self-training iterative process to help train a better classifier, is proposed. A series of experiments on both artificial and real datasets are run to evaluate the performance of our proposed framework. Experimental results clearly demonstrate that our proposed framework has better performance than some previous works in general on both artificial and real datasets, especially when the distribution of data is non-spherical. Besides, we also find that the support vector machine is particularly suitable for our proposed framework to play the role of base classifier.**

Keywords: **density peaks; self-training; semi-supervised classification; supervised learning.**

1. Introduction

Supervised learning (classification) is an active research problem in data mining and machine learning. So far, it has been widely used in power system protection, biological medicine, face recognition, image processing, and object detection, *etc* [1-6]. Supervised learning relies on the samples with class labels to train a good classifier, through which class labels can be provided for new samples. However, due to extensive expert effort along with time consumption of data labeling, it is hard to obtain sufficient labeled data. On the contrary, unlabeled data are often abundant in the real world. Consequently, having a multitude of unlabeled data and few labeled ones occurs quite often in many practical applications. In this scenario, traditional supervised learning often fails to learn an appropriate classifier with labeled data only [7]. Nevertheless, semi-supervised classification (SSC) is a learning paradigm concerned with finding a way to improve supervised learning by using unlabeled data [8-10]. Hence, in this type of learning, it is not necessary to label all the collected data for training the classifier.

Various approaches of SSC have been proposed and studied all over the world. They are usually classified depending on the different assumptions related to the link between the distribution of unlabeled and labeled data. General models are based on manifold and /or cluster assumption. If data correspond approximately to a manifold of