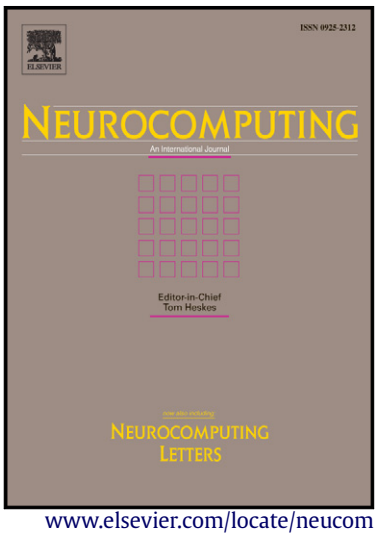

www.elsevier.com/locate/neucom

Step-wise Support Vector Machines for Classification of Overlapping Samples

Mengyu Fu, Yang Tian, Fang Wu

PII: S0925-2312(14)01699-3
DOI: http://dx.doi.org/10.1016/j.neucom.2014.12.035
Reference: NEUCOM14988

To appear in: *Neurocomputing*

Received date: 27 September 2014
Revised date: 12 November 2014
Accepted date: 14 December 2014

Cite this article as: Mengyu Fu, Yang Tian, Fang Wu, Step-wise Support Vector Machines for Classification of Overlapping Samples, *Neurocomputing,* http://dx.doi.org/10.1016/j.neucom.2014.12.035

Step-wise Support Vector Machines for Classification of Overlapping Samples

Mengyu Fu, Yang Tian, and Fang Wu

***Abstract*—Among machine learning algorithms, Support Vector Machine (SVM) is outstanding for its high efficiency and good generalization ability. This paper mainly concerns the classification performance of SVMs for multiple classes and auxiliary algorithms combined with SVMs. These auxiliary algorithms include Recursive Feature Elimination (RFE) algorithm, parameters optimizing methods and Two-Step Classification strategy. Results are given under data-based framework that classification ability and operation efficiency of SVMs are both improved when dimension is reduced; and Two-Step Classification SVM (TSC-SVM) works well under circumstances that samples overlap with each other seriously. In TSC-SVM, differences between adjacent samples are denoised by wavelet transform and magnified by a proper weighting function, after samples are sorted into correct groups in the first step. Discussions and comparisons are based on abalone dataset. According to the simulations, it is believed that step-wise SVMs have superior classification ability.**

***Index Terms*—Support Vector Machine, Genetic Algorithm, Particle Swarm Optimization, Wavelet domain denoising, Age determination, Abalone dataset**

I. Introduction

MACHINE learning is an interdisciplinary subject, involving statistics, probability theory, convex optimization and algorithm complexity analysis [1]. It is significantly essential to artificial intelligence (AI) [2]. Relevant researches are conducted with the efforts on making machines and devices mimic human's behavior, especially the learning ability. They are supposed to get new points and techniques, improving current knowledge framework to deal with problems that have never emerged before [3]. Traditional smart systems are hardly believed to be completely intelligent [4]. For their lack of learning capacity, it is not only hard for them to gain experience and self-promote when encountering mistakes, but also nearly impossible to discover and get essential knowledge on their own [5]. In other words, their reasoning is confined to deduction. It is recently found that machine learning can overcome the limitations above [6], thanks to which it has been applied to various branches of AI, including expert system, natural language understanding, computer vision, etc [7]. There are many machine learning methods, including Rote Learning (RL) [8], Learning By being Told (LBT) [9], Inductive Learning (IL) [10], Learning by Analogy (LA) [11], Explanation Based on Learning (EBL) [12], Artificial Neural Networks (ANNs) [13], Genetic Algorithm (GA) [14], Reinforcement Learning System (RLS) [15] and Support Vector Machine (SVM) [16]. With robust mathematical foundation and outstanding generalization ability, SVM shows its superiority to other algorithms, especially in classification and regression analysis [17]. This paper will pay attention to SVM's classification ability and take abalone dataset as an example to test the performance. Furthermore, relevant algorithms based on SVM and preprocessing methods are also discussed in the subsequent parts.

Based on original SVM, combined algorithms, such as SVM-RFE, are developed. In RFE arithmetic, weight vector ω is obtained from the training process, based on which corresponding sorting coefficients are calculated. According to sorting coefficients, all features can be ranked in a descending order according to their relevance to samples' classes. The greater sorting coefficient one attribute has, the more relative it is to the category label [18]. In addition, extra attentions need to be paid to the phenomenon that some classes are so close to each other that it is narrowly possible to sort them precisely [19]. In order to solve such a problem, TSC-SVM is adopted in this paper. Samples are classified into groups in the first step; then data-processing methods, including feature extractions and denoising, are applied to amplify the faint gaps between samples within a group and make it possible to realize high accuracy in the second step of classification.

The remainder of the paper proceeds as follows. Section II introduces algorithms involved, including mathematical background of SVM, normalization, wavelet domain denoising methods and optimization algorithms. Section III gives an overview of abalone dataset. Then Section IV presents comparison of parameters optimizing abilities of Grid Search (GS), Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), and results of RFE-SVM arithmetic and TSC-SVM with wavelet domain denoising. Finally, Section V provides conclusive remarks and highlighted advantages and novelties of proposed methods.

II. Algorithm

A. Support Vector Machine

1) Fundamental theory: Proposed by Cortes and Vapnik in 1995 [20], SVM has become a newly adopted method, which is mainly used in classification and regression analysis. Based on statistical learning, SVM provides an optimal design criterion for linear classifier. SVM shows its great ability not only in linear classifications but also in non-linear classifications. With the help of kernel trick, inputs are mapped onto higher dimensional feature space, making it possible

Mengyu Fu, Yang Tian, and Fang Wu are with State Key Laboratory of Robotics and System, Harbin Institute of Technology, Heilongjiang 150001, China

Corresponding Author: Fang Wu, Email: wufangzailushang@gmail.com