A Combinational Clustering Method Based on Artificial Immune System and Support Vector Machine

Zhonghua Li and Hong-Zhou Tan

Department of Electronics and Communication Engineering,
Sun Yat-sen University, Guangzhou 510275, China
{lizhongh,issthz}@mail.sysu.edu.cn

Abstract. Clustering is one branch of unsupervised machine learning theory, which has a wide variety of applications in pattern recognition, image processing, economics, document categorization, web mining, etc. Today, we constantly face how to handle a large number of similar data items, which drives many researchers to contribute themselves to this field. Support vector machine provides a new pathway for clustering, however, it behaves bad in handling massive data. As an emergent theory, artificial immune system can effectively recognize antigens and produce the memory antibodies. This mechanism is constantly used to achieve representative or feature data from raw data. A combinational clustering method is proposed in this paper based on artificial immune system and support vector machine. Experimentation in functionality and performance is done in detail. Finally a more challenging application in elevator industry is conducted. The results strongly indicate that this combinational clustering in this paper is of feasibility and of practice.

Indexed Terms: Data clustering, combinational clustering, artificial immune system, support vector machine.

1 Introduction

For a simpler approach to identifying similar data items, *clustering analysis* may be used. It refers to a large number of algorithms and methods for grouping together similar data items with the groupings relationships between the items. It is a useful first step in the data analysis process because it can guide the efforts of any further study that might required. Clustering analysis has been widely applied to such fields as pattern recognition, image processing, economics, document categorization, web mining, etc [1]. A common question facing researchers in many areas is how to organize observed data into meaningful structures. The general clustering methods include Tree Clustering, Block Clustering, K-means Clustering, Evolutionary Maximization Clustering, and so on [3].

Since 2000, *support vector machine* (SVM) have gathered numerous eyeballs around the machine-learning world, which almost includes three branches: classification, regression and clustering [2]. Unlike support vector classification, support vector clustering is unsupervised learning. Data points are mapped from data

B. Gabrys, R.J. Howlett, and L.C. Jain (Eds.): KES 2006, Part I, LNAI 4251, pp. 153 – 162, 2006.

space to a high dimensional feature space by means of a Gaussian kernel. In feature space, the smallest sphere is searched which can enclose the image of the data. This sphere is then mapped back into data space. Several contours are produced to enclose the data points. Points enclosed by each separate contour are of the same cluster. However, large-scale data items from data space are directly mapped to a high dimensional feature space, which increases the time complexity of SVM. To overcome this demerit, it is essential to extract the feature data from the raw dataset, which will be feed to SVM for clustering analysis.

In the recent years, artificial immune system interests any researches due to its rich immune mechanisms and powerful information processing capability [4-7], where de Castro and von Zuben creatively applied *artificial immune system* (AIS) to data clustering [6]. A classical *artificial immune network* called aiNet was constructed under the spirit of biological immune system, here the raw data was considered as the antigens of immune system, and the feature data was outputted as memory cells through the evolutionary interaction between the given antigens and a number of antibody candidates. Later an improved clustering method based on AIS was presented, which explores a new pathway for clustering large-scale data items. As long as the related parameters are chose properly, AIS is no longer time-consuming for any type of data items.

In the past, we attempted to combine AIS and SVM for data clustering, where AIS is responsible for extracting feature data from the raw data and SVM focuses on clustering the extracted feature data [12]. Based on the previous work, this paper will propose an improved combinational clustering method (ICCM) for further study. Our goal is advance ICCM more efficient and more stable. In the entry of this combined method, we will present a tactful policy that adds an extra block for data processing from the viewpoint of systems solutions.

The remainder of this paper is organized as follows. Some basic features about biological immune system and SVM are firstly reviewed in Section 2. Section 3 proposes a modified combined clustering method based on AIS and SVM, and gives some technical details. Simulation experiments and discusses are listed in Section 4. Finally conclusions are drawn in Section 5.

2 Features of Related Theories

A. Support Vector Machine

SVMs have been widely adopted for classification, regression and novelty detection, recently used for clustering analysis too. The basis of this support vector clustering is density estimation through SVM training. Support vector clustering is a boundary-based clustering method, where the support information is used to construct cluster boundaries. It is able to detect arbitrary shape clusters with a hierarchical structure in high dimensional data. It provides a way to deal with outliers and by using kernel methods; explicit calculations in feature space are not necessary [2].

Given $\{x_i\} \in \chi$ be a dataset with N points, Φ is a non-linear transformation from χ to some high dimension feature space. The smallest enclosing sphere of radius R is looked for. When soft constraints are incorporated by adding a slack variable ξ_j, R should satisfy the following constraint.