
IGI PUBLISHING **ITB15223**
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication, **Data Mining and Knowledge Discovery Technologies** edited by D. Taniar © 2008, IGI Global

Chapter VI

K-Means Clustering Adopting rbf-Kernel

ABM Shawkat Ali, Central Queensland University, Australia

Abstract

Clustering technique in data mining has received a significant amount of attention from the machine learning community in the last few years as one of the fundamental research areas. Among the vast range of clustering algorithms, K-means is one of the most popular clustering algorithm. In this research, we extend the K-means algorithm by adding well known radial basis function (rbf) kernel and find better performance than classical K-means algorithm. It is a critical issue for rbf kernel; how can we select a unique parameter for optimum clustering task. This chapter will provide a statistical-based solution on this issue. The best parameter selection is considered on the basis of prior information of the data by the maximum likelihood (ML) method and nelder-mead (N-M) simplex method. A rule based meta-learning approach is then proposed for automatic rbf kernel parameter selection. We consider 112 supervised data set and measure the statistical data characteristics using basic statistics, central tendency measure, and entropy-based approach. We split these data characteristics using the well-known decision tree approach to generate the rules. Finally, we use the generated rules to select the unique parameter value for rbf kernel and then adopt in K-means algorithm. The experiment has been demonstrated with 112 problems and 10 fold cross validation methods. Finally, the proposed algorithm can solve any clustering task very quickly with optimum performance.

Introduction

Data mining can quite often be defined as a useful hidden knowledge extraction process from a huge database. Basically, two types of techniques, supervised and unsupervised, are using to extract this knowledge. When the problem is not predefined, then the researcher always chooses the unsupervised technique to solve their problems. Now a days, a good number of unsupervised techniques introduced by researchers and are free for use. K-means algorithm is an old unsupervised technique but still it is a popular technique. The job of unsupervised technique is called clustering. In 1967, MacQueen (1967) developed the K-means clustering algorithm for classification and analysis of multivariate observations. Since then, while unsupervised techniques have been studied extensively in the areas of statistics, machine learning, and data mining (Zalane, 2007), the K-means algorithm has been applied to many problem domains, including the area of data mining, and has become one of the most used clustering algorithms. Matteucci (2007) even said that the K-means algorithm is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. Recently we found after adding the kernel (Vapnik, 1995) components with K-means algorithm, it became a more popular and powerful unsupervised technique (Dhillon, Guan, & Kulis, 2004, 2005; Kulis, Basu, Dhillon, & Mooney, 2005; Zhang & Rudnicky, 2002). In general, kernel function implicitly defines a non-linear transformation that maps the data from their original space to a high dimensional feature space where the data are expected to be more separable. As a result, the kernel methods may achieve better performance by working in the new space (Zhang et al., 2002). Kernel method is comfortable for both linear and non-linear space. Moreover, it can handle any high dimensional data in their transformation space. Three types of classical kernels namely linear, polynomial, and rbf are introduced initially with kernel-based learning algorithms. Among these, rbf kernel is quite popular and many popular kernels for a specific problem are available today (Cheng, Saigo, & Baldiet, 2006; Ou, Chen, Hwang, & Oyang, 2003). The critical issue of rbf kernel is to select unique parameter within a range of values. We proposed a rule-based methodology for rbf kernel parameter selection using statistical data characteristics (Ali & Smith, 2005). In this research, we introduced this method with classical K-means algorithm. First, we do clustering 112 supervised problems by K-means algorithm. The data are considering from two different sources (Blake & Merz, 2002; Lim, 2002). After that, we implemented the rbf kernel with automated parameter selection in K-means algorithm and perform the clustering task with the similar data.

This chapter is organized as follows: In the next section we shall provide the theoretical frameworks regarding K-means clustering, rbf kernel, and it's automated parameter selection with statistical formulation and measures. Then we shall describe the analyses of the experimental results. Finally, we conclude our research toward the end of this chapter.