Pattern Recognition 74 (2018) 185–197

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Learning representative exemplars using one-class Gaussian process regression

Youngdoo Son[a], Sujee Lee[b], Saerom Park[b], Jaewook Lee[b,*]

[a] Department of Industrial and Systems Engineering, Dongguk University-Seoul, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, South Korea
[b] Department of Industrial Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea

ARTICLE INFO

Article history:
Received 23 August 2016
Revised 27 June 2017
Accepted 2 September 2017
Available online 11 September 2017

Keywords:
Representative exemplars
One class Gaussian process regression
Support-based clustering
Automatic relevance determination
Kernel methods

ABSTRACT

An exemplar is an observation that represents a group of similar observations. Exemplars from data are examined to divide entire heterogeneous data into several homogeneous subgroups, wherein each subgroup is represented by an exemplar. With its inherent sparsity, an exemplar-based learning model provides a parsimonious model to represent or cluster large-scale data. A novel exemplar learning method using one-class Gaussian process (GP) regression is proposed in this study. The proposed method constructs data distribution support from one-class GP regression using automatic relevance determination prior and heterogeneous GP noise. Exemplars that correspond to the basis vectors of the constructed support function are then automatically located during the training process. The proposed method is applied to various data sets to examine its operability, characteristics of data representation, and cluster analysis. The exemplars of some real data generated by the proposed method are also reported.

1. Introduction

Finding representative points or exemplars among similar observations has been extensively studied to identify points that can represent similar groups of observations, as well as clusters of data points that are robust to noise and outliers. *k*-medoids clustering [17] is an early approach of clustering method that uses medians as representative points instead of the means of *k*-means clustering. Numerous studies examined *k*-medoids clustering method and its variants. Affinity propagation (AP) [10] has received considerable attention for a decade. Compared with other exemplar-based clustering algorithms, such as *k*-medoids clustering, AP identifies representative points using the message passing function of the graph and finds reasonable clusters. Moreover, AP is robust to initialization. However, both algorithms focus on finding the convex clusters of data points with only one representative point for each cluster. These algorithms cannot constantly represent exemplars for non-convex shape data distribution.

Support-based clustering methods with kernels that originate from support vector clustering (SVC) [5] are another set of new methods. These methods construct a function that estimates the support of a given data distribution and divides the data set into several similar groups based on the level set of support function. These methods were successfully applied in solving difficult clustering problems because of their better ability to detect complicated non-convex shapes than traditional clustering methods. Original SVC or its variants [5,33,44] use the complete or partial graph of the given data points to assign cluster labels. These methods do not directly locate the representative points because they focus only on cluster labeling that utilizes information of the support vectors found at cluster boundaries. By contrast, equilibrium-based approaches [14,23,24] find equilibrium vectors, wherein each instance converges under the associated dynamic system. Equilibrium vectors can represent converging instances with the same cluster label. However, these vectors are not exemplars because they do not belong to the original data set. Moreover, support-based methods are memory intensive and computationally expensive because the models re-evaluate most of the data samples at every simulation.

A novel algorithm that learns representative exemplars is proposed in this study to overcome the bottlenecks of exemplar- and support-based methods. The proposed method automatically finds the representative exemplars of given data points and constructs the sparse support function centered on the exemplars. A number of strategies are provided to assign the cluster labels to input points using the merits of the obtained exemplar and sup-

* Corresponding author.
E-mail addresses: youngdoo@dongguk.edu (Y. Son), sujee0524@snu.ac.kr (S. Lee), psr6275@snu.ac.kr (S. Park), jaewook@snu.ac.kr (J. Lee).

http://dx.doi.org/10.1016/j.patcog.2017.09.002
0031-3203/

port function. First, we employ the one-class Gaussian process (GP) regression model with a variance function that depends only on training input data. We use automatic relevance determination (ARD) prior distribution to enable the model to automatically identify the representative points of input data. ARD prior distribution leads to variance functions in a dense area with large values and small values in a sparse area; these variance functions are the opposite of those observed in the traditional GP regression model [18,35]. We also introduce a variable GP noise instead of the traditional constant Gaussian noise to make the GP regression model compatible with the GP classification model for one-class problem. This approach allows us to obtain a likelihood function that builds a sparse model represented by exemplars. Thus, the proposed method constructs the kernel-based support function represented by a small number of representative data points that function as exemplars in a grouped cluster. The proposed method then finds non-convex clusters with exemplars because it shares the benefits of exemplar- and support-based methods.

The remainder of this paper is organized as follows. Previous studies related to the proposed method are briefly described in the following section. Section 3 presents the proposed method used to find representative exemplars for unlabeled points through one-class GP regression model. The implementation strategy used to determine the hyper-function and hyper-parameters is discussed in detail. We also show that the proposed method can theoretically estimate the support of an unknown data distribution using generalization error bound to ensure that the method can be sued to find data clusters. Section 4 shows the experimental results applied to several kinds of data sets for data representation and clustering. Section 5 concludes study with some discussions.

2. Related work

This section briefly reviews the recent results and milestone literature of the topics related to the proposed method.

k-medoids clustering is a traditional and popular exemplar-based clustering algorithm [17]. Given an initial set of representative points, *k*-medoids clustering assigns the other points to the nearest representative points and constructs the clusters. A new set of representative points is identified by selecting the medians of the obtained clusters. Finally, clusters are selected by repeating the two steps until convergence is reached. *k*-medoids clustering is robust to outliers because it uses the medians of groups instead of the means that are often deteriorated by outlying values. The former selects the existing instance, namely, the median, as the representative point of a cluster. By contrast, the latter employs a point without the existing instance. However, *k*-medoids clustering has several disadvantages that can also be found in *k*-means clustering. First, *k*-medoids clustering finds clusters with convex shapes. Thus, data points with sophisticated distribution cannot be accurately clustered. Second, the number of clusters *k* should be provided before the algorithm is conducted. The domain knowledge or the validation method can be employed to determine the number of clusters. Third, *k*-medoids clustering is sensitive to initialization. The different initial points can occur in different clusters. Finally, the computation used to find the clusters based on the median points is not trivial. *k*-medoids clustering was improved in several ways to address these limitations. Park and Jun [32] suggested the fast algorithm for *k*-medoids clustering to reduce computational time through one-time calculation of the distance matrix. Yang and Zhang [45] proposed the kernel *k*-medoids clustering to find non-convex shaped clusters of uncertain data points by defining the distances between two points by those in high-dimensional mapped feature space. Krishnapuram et al. [21] suggested the fuzzy version of the k-medoids algorithm to find soft clusters instead of hard ones.

AP [10] is a similarity-based clustering algorithm that finds the representative points through message passing of given data points. AP computes the "responsibility" $r(i, k)$ that measures how well a point k represents the exemplar for a point i and the "availability" $a(i, k)$ that measures how appropriate it would be for a point k to be an exemplar for a point i. The responsibilities and availabilities are then iteratively and alternatively updated until the network becomes stable. Exemplars are data points with positive "responsibility + availability." AP received significant attention because it solves some of the limitations mentioned above. AP does not require the pre-determined number of clusters and is not sensitive to initialization. In addition, AP finds clusters in reasonable times with the complexity $O(N^2T)$, where N is the number of data points and T is the number of iteration. Thus, AP has been applied to several real problems, including forest fragmentation analysis [36], gene selection problem [8], and facility location problem [7]. However, AP has some drawbacks. First, AP assumes the rather simple shapes of clusters as one cluster that contains one exemplar. Second, AP requires several parameters, such as preference value for every point, damping factors, and maximum number of iterations. Different choices of parameters can result in different clustering findings, including the number of clusters. Finally, computation cost may cause a problem for large data sets. Wang et al. [42] proposed multi-exemplar AP to identify appropriate clusters for sophisticatedly distributed instances. In multi-exemplar AP, the complex shape cluster can be constructed by enabling the clusters to contain several exemplars. However, given that the optimization of multi-exemplar AP is NP-hard, Wang et al. [42] suggested another belief propagation that finds the approximate solution; by contrast, the proposed method finds the multi-exemplar clusters in polynomial computations. A number of AP variants were also proposed to improve the speed of propagation. Jia et al. [13] proposed a variant of AP that could be applied to the sparse similarity matrix to reduce the computation time. Shang et al. [39] developed fast AP through the multi-level approach that considers both local and global structure information.

Support-based clustering methods are composed of two stages. The first stage involves construction of support function that detects the cluster structure of a given data distribution. A widely used support function is constructed from the support vector domain description algorithm [5] or the one-class support vector machine algorithm. Another function is obtained from the variance of the one-value GP regression using squared exponential kernel with Gaussian noise [18]. According to the algorithms employed for the construction of the support function, the former is called SVC, whereas the latter is called Gaussian process clustering (GPC). The second stage assigns the cluster labels of the data points by determining whether two data points belong to the same cluster or not using the level set of the constructed support function. Several studies attempted to reduce labeling time. Graph-based labeling uses the complete or partial graph of the given data points to assign their cluster labels, but this graph-based approaches usually require excessive time to find the cluster labels. These methods include fully connected graph approach [5], approximated graph technique [44], spectral graph partitioning [33], ensemble combination [34], chunking strategy [3], pseudo-hierarchical technique [11], and cone cluster labeling [26]. Equilibrium-based labeling finds the converging point of each instance along the support function and assigns the label of those converging points to reduce labeling time. However, solving the dynamic systems for each point to find its converging point requires a significant amount of time. These equilibrium-based labeling includes dynamic system-based