Pattern Recognition 48 (2015) 3673–3687

Contents lists available at ScienceDirect

Pattern Recognition

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

A novel validity index with dynamic cut-off for determining true clusters

M.S. Bhargavi *, Sahana D. Gowda

Department of Computer Science & Engineering, BNM Institute of Technology, Bengaluru 560070, Karnataka, India

ARTICLE INFO

Article history:
Received 30 August 2014
Received in revised form
25 March 2015
Accepted 19 April 2015
Available online 29 April 2015

Keywords:
Clustering
Validity index
Dynamic cut-off
True clusters
k-means clustering
Hierarchical agglomerative clustering

ABSTRACT

In a multi-surveillance environment, voluminous data is generated over a period of time. Data analysis for summarization and conclusion has paved a way for the need of an efficient clusterization. Clustering, an unsupervised way of learning about data aims at defining clusters. Validation of clusters formed indicates the trueness of the clusters. In this paper, a novel validation technique with dynamic termination of clustering process has been proposed to obtain true clusters. In the validation process, the validity index is based on both global cluster proximity relationship and local proximity relationship. The validity index is computed for validating the available clusters using 'within-cluster sum-of-squares', 'between-cluster sum-of-squares', 'total-sum-of-squares', 'intra-cluster distances' and 'inter-cluster distances'. The ratio between two consecutive validity indices is the extent of variation which specifies the cut-off point. Cut-off terminates the clustering process dynamically indicating the number of clusters and validates the obtained clusters. The proposed method is tested on several real and synthetic data sets. Comparisons with the existing methods demonstrate the efficiency of the proposed method in detecting true clusters.

1. Introduction

Voluminous data generated in real-time poses great challenges for its analysis. Based on available knowledge, data can be analyzed either in a supervised or in an unsupervised manner, with classification and prediction being feasible approaches for the former and clustering a well-known approach for the latter.

Clustering aims at grouping data into coherent clusters based on the proximity of samples in an *n*-dimensional space [1–3] The cohesive strength of samples within each cluster and the distinctness of the clusters need to be assessed to determine the trueness of these clusters. The process of evaluating the trueness of the formed clusters is known as cluster validation [4–6]. This validation process estimates the occurrence of true clusters by a validity index [7].

Validation process has been categorized into local and global validation methods [8]. Global methods compute the proximity relationship among the clusters in terms of validity index for the obtained clusters. Assessment of every cluster combination obtained, in order to draw conclusions regarding optimal clusters, makes the model computationally complex.

Local cluster validation methods possess criteria with parametric assumptions, which terminate the clustering process dynamically, at the formation of true clusters. This reduces the computational complexity as the need for further clustering and validation, once true clusters have been obtained, is eliminated. The local methods assess trueness based on optimal local proximity in terms of merging or splitting clusters only and does not consider the proximity relationship among other clusters.

This paper aims to develop an effective validation process that terminates the clustering process dynamically, at the formation of true clusters. The validation process is a hybrid approach that computes the validity index based on the proximity relationship between clusters (global proximity) and the proximity relationship within each cluster, which allow these clusters to merge at subsequent level (local proximity). This ensures that validation determines globally optimal cluster memberships along with local detection of cluster stability by eliminating superfluous computations.

Clustering and validation are performed for cluster combinations starting with k_{max} (maximum cluster combinations) and moving towards k_{min} (minimum cluster combinations). The validity index and the cut-off ratio are computed between consecutive levels. To determine the validity index, 'within-cluster sum-of-squares', 'between-cluster sum-of-squares', 'total sum-of-squares',

* Correspondence to: BNM Institute of Technology, Post Box no. 7087, Department of Computer Science & Engineering, 27th cross, 12th main, Banashankari stage II, Bengaluru 560070, Karnataka, India. Tel.: +91 9845643435.
E-mail address: ms.bhargavi@gmail.com (M.S. Bhargavi).

http://dx.doi.org/10.1016/j.patcog.2015.04.023
0031-3203/

'intra-cluster distances' and 'inter-cluster distances', are computed. A dynamic cut-off criterion is attained by finding the ratio between the current validity index and the predecessor validity index. This ratio determines the extent of variation between the current and predecessor level. If the extent of variation is not high, the ratio tends to be closer to one. This indicates that the current cluster solution is acceptable and the process progresses to the next level. If the extent of variation is very high, the ratio tends to be closer to zero. In such situation, the process terminates indicating that the current cluster solution is unacceptable and the true clusters have been obtained in the predecessor level.

Experimental results on real and synthetic data sets of varying complexities, demonstrate the effectiveness of the method in determining true clusters. The competency of the proposed method is proved by comparing it with well-known cluster validity indices available in literature.

The rest of the paper is organized as follows. Section 2 focuses on the state-of-art. Section 3 presents a detailed description of the proposed methodology. Section 4 depicts experimental analysis and results. Section 5 deals with comparative analysis along with time complexity analysis. Section 6 presents the conclusions.

2. State-of-art

Several methods have been proposed by various authors to determine true clusters. These methods are based on validating the obtained clusters, which can be categorized into methods based on model-fitting [9–13], stability-based methods [14–21], and methods based on cluster validity indices [4–7, 22–26].

Likelihood estimation [9–13] is a method based on 'model-fitting' in which, models are designed to best fit the data, based on information criteria such as Bayesian information criterion [27], Akaike information criterion [28], Deviance information criterion [29] or subspace information criterion [30]. These methods require strong parametric assumptions and are computationally expensive.

Resampling [14–18] and combining the results of multiple clustering [19–21] are two different approaches to stability-based methods. Resampling methods cluster multiple combinations of data obtained by random sub-sampling, random splitting and bootstrapping, until these clusters meet the stability criteria. In methods 'combining the results of multiple clustering', true clusters are assessed using the results obtained by running the same algorithm at different settings or by running different algorithms at the same settings. Stability-based methods are computationally expensive, as it requires the clustering to be rerun many times either on different combinations of data or different clustering settings.

Based on the computation of validity indices for the clustered data, two different approaches have been proposed. One is the 'graph evaluation' method [22–24], which is a post-processing method and the 'run-time evaluation' method [4,5], which determines true clusters automatically. Graph evaluation method is also known as 'locating the knee of an error curve' method [22] or the 'elbow' method [23,24]. This method uses an evaluation graph, which is a plot of the number of clusters versus the measure of the evaluation. The point on the curve that either maximizes or minimizes the measure of evaluation being used is chosen as the correct cluster number.

Run-time evaluation methods are based on external and internal criteria [4,25]. Methods that use external criterion for validation utilize previous knowledge of the data, while those that use internal criterion are based on the information intrinsic to the data alone. Validation with internal criterion is classified into global and local cluster validation [8]. A detailed account of various validity indices is given by Desgraupes [7] and Arbelaitz et al. [26]. Arbelaitz et al. provide an extensive comparative study to evaluate the effectiveness of 30 different cluster validity indices in determining true clusters.

Methods such as likelihood estimation, resampling, and combining the results of multiple clustering, require either strong parametric assumptions, or a clustering algorithm to be rerun several times. Global cluster validation methods run the validation for every potential cluster combination and do not automatically stop the clustering process once true clusters are obtained. The methods that do have stopping criteria are all local methods, which require parametric assumptions for automatically terminating the clustering process. Moreover, these methods assess only a part of the data for computing the cut-off point. It is evident from the survey that the available cluster validation methods lack dynamic termination without parametric assumptions and global optimization. This paper proposes a novel, hybrid cluster validity measure with dynamic cut-off for cluster validation.

3. Proposed methodology

Data samples from an unsupervised environment are unlabeled observations of a scenario in different dimensions. Clustering algorithms are used to find groups among unlabeled observations that are further analyzed for knowledge extraction. At the high end of the spectrum of clustering algorithms are two major categories: partitional and hierarchical clustering [1,9,31]. Partitional clustering algorithms minimize a given clustering criterion by iteratively relocating the data points between clusters until an optimal partition is attained [3]. Hierarchical clustering starts from a single cluster and ends with multiple clusters or vice versa. The clustering technique that starts from single cluster and splits into many in a hierarchical fashion is known as 'divisive hierarchical clustering', which is a top-down approach [2]. The technique that starts from many single clusters and merges them hierarchically is known as 'agglomerative hierarchical clustering', which is a bottom-up approach [2].

In partitional clustering, the range of cluster numbers to be assessed is decided a priori. The optimum range of cluster numbers chosen for analysis could be based on Bezdek's suggestion of $k_{min}=2$ to $k_{max}=\sqrt{n}$ [34], where n is the number of samples in the data set. This is considered as thumb rule in clustering literature and the validition to choosing the best range from $\sqrt{n}$ is out of the scope of this research work. Clustering and validation using the proposed method starts from the maximum cluster number towards the minimum, in the specified range. For every cluster combination obtained, a test of optimality is performed by computing the validity index and the cut-off ratio between the current and the predecessor level. If an intermediate cluster combination is voted as the best based on the cut-off ratio, the process terminates automatically.

In hierarchical agglomerative clustering, validation is performed for clusters obtained over every merge at the corresponding level. At every level, validity index and the cut-off ratio between the current and the predecessor level is computed. If the current merge is unacceptable based on the cut-off ratio, the predecessor level is voted as the best and the process terminates automatically.

For a better understanding of the proposed method, the following notations are used throughout the description:

Xp	data sample at pth observation in m dimensional feature space
n	number of m dimensional samples in the dataset
m	number of features