Available online at www.sciencedirect.com

ScienceDirect

Karbala International Journal of Modern Science 1 (2015) 200–211
http://www.journals.elsevier.com/karbala-international-journal-of-modern-science/

An integrated robust semi-supervised framework for improving cluster reliability using ensemble method for heterogeneous datasets

Smita Prava Mishra [a,]*, Debahuti Mishra [b], Srikanta Patnaik [b]

[a] *Dept. of Computer Sc. & Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India*

[b] *Dept. of Computer Sc. & Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India*

Received 28 September 2015; revised 9 November 2015; accepted 9 November 2015
Available online 18 December 2015

Abstract

Data mining literature offer some clustering techniques. But when we implement even an effective clustering technique, the results are found unreliable. The efficacy of the technique come under scrutiny. Here, the proposal is about an integrated framework, which ensures the reliability of the class labels assigned to a dataset whose class labels are unknown. The model uses PSO-*k*-means, *k*-medoids, *c*-means and Expectation Maximization for data clustering. This model integrates their results through majority voting cluster ensemble technique to enhance reliability. The reliable outcomes serve as the training set for the classification process through Bayesian classifier, Multi Layer Perceptron, Support Vector Machine and Decision tree. The predicted class labels by majority of classifiers through bagging classifier ensemble method are included with the training set and in combination, designated as the set with known class labels. Heterogeneous datasets with unknown class labels but known number of classes, after being treated through this model would be able to find the class labels for a significant portion of the data and may be accepted with reliability. The evaluation procedure has been performed by following the Dunn's, Davies–Bouldin and Modified Goodman–Kruskal indexing techniques for internal validation and probabilistic measures such as Normalized Mutual Information, Normalized Variation of Information and Adjusted Random Index which are appropriate measures of goodness-of-fit and robustness of the final clusters. The predictive capacity of the model is also validated through probabilistic measures and external indexing techniques such as Purity Measure, Random Index and F-measure.

Keywords: Data clustering; Cluster ensemble; Internal indexing; External indexing; Probabilistic measures

1. Introduction

Dealing with unclassified data is a real time challenge for data miners. After years of research on several clustering algorithms, researchers could not succeed in designing a standalone robust clustering algorithm for heterogeneous datasets which could

* Corresponding author.
E-mail addresses: smitaprava@gmail.com (S.P. Mishra), mishradebahuti@gmail.com (D. Mishra), prof.srikantapatnaik@gmail.com (S. Patnaik).

Peer review under responsibility of University of Kerbala.

http://dx.doi.org/10.1016/j.kijoms.2015.11.004

assign reliable class labels to unclassified data. With the aim to improve reliability and robustness of the clustering outcomes, this paper proposes a semi-supervised method for clustering data where the class labels of the data are unknown. Multiple clustering techniques [1−4] such as PSO-*k*-means [5,6], *k*-medoids [7], *c*-means [8] and Expectation Maximization (EM) [4] are applied on datasets of diverse domains. To improvise the reliability of the clustering results, the majority voting, otherwise known as bagging [2,9,10] cluster ensemble technique is adopted. Through the voting method, each dataset is segregated into two partitions. One having pure majority upon the obtained clustering results and the other data partition, without pure majority. Subsequently, a learning environment is simulated with multiple classifiers such as Bayesian classifier [11−14], Multi Layer Perceptron (MLP) [13−15], Support Vector Machine (SVM) [13−15] and Decision tree [13−15] classifiers being individually trained with the data partition with pure majority where label obtained from agreed clustering techniques is treated as the class label of the training set. After training, the classifiers are tested with the remaining partition of the data without pure majority. The testing results of multiple classifiers are again ensembled through majority voting [2]. The remaining data without pure majority after ensemble is discarded and rest of the data is accepted with their class labels. The evaluation procedure is performed to verify reliability of the results through various validation methods such as internal indexing techniques, external indexing techniques and statistical methods like probabilistic measures for clustering results. Our experimentation includes: (a) internal indexing techniques [16−19] such as Dunn's index [20], Davies−Bouldin index [21] and Modified Goodman−Kruskal ($GK_{modified}$) index [22] which does not take any reference of the known class labels and only consider tightness of the intra-cluster elements and separation among inter-cluster elements for measuring the quality of the clusters; and (b) the probabilistic measures [23−25] taken for clustering result validation are Normalized Mutual Information (NMI), Normalized Variation of Information (NVI) and Adjusted Random Index (ARI) which relies upon statistical methods for measuring overlapping of comparative classes. The above two validation strategies are appropriate measures of goodness-of-fit and robustness of the final clusters. The external indexing techniques [18,19] applied for validation with reference to the real class labels of the dataset are Purity Measure [26], Random Index [27] and F-measure [26]. They are supervised methods of result validation and exploit the known information about a dataset for comparison purpose. The predictive capacity of a model is validated through probabilistic measures and external indexing techniques.

After treatment through the model, the class labels obtained for a significant partitions of the datasets can be accepted as reliable with credible class labels. As the framework relies both on unsupervised as well as supervised methods, it may be designated as a semi-supervised method of data clustering.

The article is structured as follows: first, the schematic description section shows the layout of the proposed integrated semi-supervised framework for class label determination for heterogeneous datasets. Second, the method selection and parameter discussion is presented which also highlights the clustering, classification, internal and external indexing mechanisms along with probabilistic measures with reasoning and justifications. In the experimentation section, the description of datasets along with stepwise empirical evaluation is discussed. The result analysis section critically evaluates the significance of the findings described in experimentation section. Finally, the conclusions of this work are summarized and future directions are highlighted.

2. Schematic description

Fig. 1 describes the schematic representation of the environment simulated for the work. It also identifies the datamining and validation techniques used for the same. Datasets with removed class labels are taken at first for the purpose of clustering. Multiple clustering techniques are applied on those datasets individually. The clustering results are then integrated through cluster ensemble technique on per tuple basis. Then based on the results, each dataset is segregated into the training set with majority agreed consensus cluster determined and testing set whose class/cluster labels are not yet known. Then, each training set is used to train multiple classifiers. The testing sets are now given to the classifiers for identification of their class/cluster labels. Again the consensus is taken to obtain a single class label for each tuple. The training tuples and the tuples with consensus in classification techniques are designated as the final dataset with known class/cluster labels. The remaining tuples with still ambiguity in their class determination are discarded from the dataset.

The final set of tuples with their class labels are then treated for verification and acceptability of the results