Compte Rendu d'Articles HMIN326

2020 - 2021

Context-based Distance Learning for Categorical Data Clustering

Amir SHIRALI POUR
Student Number: 21912984
Konstantin TODOROV

14 December 2020

1. Abstract

Data clustering enables to group attributes into groups sharing similarities and calculating the distance between the pair of these numeric attributes. However, calculating the difference between categorically described numerical attributes is not possible. In the literature, Multiple studies have suggested different approaches to the clustering process of categorical attributes such as K-MODES where unlabeled data points and clusters are represented with distinguished points allowing to calculate the overlap distance. The ROCK approach uses links between data points including the similarity between them, this hierarchical approach has a cubic complexity related to the dataset size, which makes it disadvantageous when applied to a large set. Another hierarchical approach is LIMBO which is a slower algorithm based on building and containing cluster information on Distributional Cluster Features trees. Other approaches exist that include clustering based on graph and hyper-graph partitioning such as CLICKS. In this study, the proposed solution to the lack of efficiency and time complexity associated with large dataset is the DILICA method (Distance Learning in Categorical Attributes). It defines a context from an informative attributes subset and uses it to compute the distance between values of the same attribute. The study shows that combining the DILCA method with partitional and hierarchical methods (KMODS and Ward hierarchical clustering: HCL) and comparing them to ROCK and LIMBO methods demonstrates the superiority of DILCA in average accuracy and normalized mutual information achieved in clustering, as well as in a minimal computing time performance. The results demonstrate that the DILCA solution is effective in categorical data and distance computation and nominate it to be a strong general candidate to use in the data mining and clustering process.

Answer 2

1. selection of a relevant subset of the whole attributes set that is used as the context for a given attribute;

2. computation of the distance measure between pair of values of the same attribute using the context defined in the previous step.

Its goal is to select a subset of relevant and not redundant features and discard all the other ones w.r.t. a given class attribute.

Answer 3

For the evaluation of our distance learning approach on categorical data, two collections of datasets are used. The first collection consists in real world data sets downloaded from the UCI Machine Learning Repository and the second collection contains synthetic datasets produced by a data generator using Gaussian distributed random attributes. Notice that Breast-w and Sonar contain numerical variables. Indeed, they have been discretized using the supervised method proposed in.

Answer 4

To evaluate the impact of σ parameter Mushroom and Votes datasets has been used. For each dataset the behavior of K-MODES$_{\text{DILCA}}$ is plotted. We let vary the parameter σ from 0 to 1 with steps of 0.1. When the parameter is equal to 0 all the features are included in the context. it is observed that the influence of different settings of σ w.r.t. accuracy is small. In both datasets, the variation in accuracy is very low (less than 0.50%).

Although there is no general law about how to choose this parameter, we estimate that its impact is less important than standard clustering parameters.

Answer 5

The components K-MODES$_{\text{DILCA}}$, HCL$_{\text{DILCA}}$, ROCK and LIMBO has been analyzed.

In almost all the experiments, our approach achieves the best results in at least one category of clustering, and in some cases (Sonar and Votes), the performance parameters are sensibly better than in ROCK and LIMBO. The only exception is SynA, where NMI computed for ROCK is slightly higher than NMI achieved by HCLDILCA. However, the Accuracy measured for ROCK is lower than the one achieved by HCLDILCA. Moreover, we observed that ROCK is very sensitive to the parameter value, and in many cases this algorithm produces one giant cluster that includes instances from more classes.