Adaptive ensembling of semi-supervised clustering solutions

Abstract:

Semi-supervised clustering is an important sub-field of clustering and is widely applied in different areas, such as image processing , multimedia, pattern recognition and bioinformatics. For example, Biswas et al. applied the constrained clustering algorithm for image analysis. Liu et al. proposed a novel semi-supervised matrix decomposition approach, and applied it to image processing and document clustering. Lai et al. proposed a new interactive semi-supervised clustering model for large image database indexing. He et al. proposed multi-level random walk based semi supervised clustering. Jiao et al. designed fast semi-supervised clustering.

Existing System:

We focus on constrained clustering, which belongs to the class of semi-supervised clustering approaches. Constrained clustering integrates a set of must-link constraints and cannot-link constraints into the clustering process. The must-link constraint means that two data samples should belong to the same cluster, while the cannot-link constraint means that two data samples cannot be assigned to the same cluster. Traditional constrained clustering approaches have two limitations: (1) They do not consider how to make full use of must-link constraints and cannot-link constraints; (2) Some methods do not take into account how to deal with high dimensional data with noise.

Disadvatages:

- They do not consider how to make full use of must-link constraints and cannot-link constraints.
- Some methods do not take into account how to deal with high dimensional data with noise.

Proposed System:

An adaptive semi-supervised clustering ensemble framework (A-RSEMICE) for high dimensional

data clustering. When compared with traditional semisupervised clustering approaches, A-RSEMICE is characterized with the following three properties: (1) A newly proposed transitive closure based constraint propagation approach is adopted to make use of the transitive closure operator and the constraint propagation to fully explore how to use all useful must-link and cannot-link constraints. (2) A-RSEMICE adopts the random subspace based semi-supervised clustering ensemble framework to integrate the clustering solutions obtained by different transitive closure operators from multiple datasets into a unified clustering solution. (3) A newly designed adaptive process is adopted to search for the optimal subspace set. We performed a thorough analysis of the properties of A-RSEMICE in the experiments, and draw conclusions as follows: (1) The transitive closure operator and the confidence factor each plays an important role in attaining good performance for A-RSEMICE. (2) The adaptive process is useful for A-RSEMICE to obtain better results. (3) ARSEMICE outperforms most of the state-of-the-art approaches when dealing with high dimensional cancer datasets. In the future, we will adopt a suitable cluster validity index to determine the number of clusters.

Modules:

- Transitive closure based constraint propagation clustering approach.
- Random subspace based semi supervised cluster ensemble framework.
- Adaptive semi-supervised cluster ensemble framework.

SYSTEM REQUIREMENTS

H/W System Configuration:-

Processor	-	Pentium –III
RAM	-	256 MB (min)
Hard Disk	-	20 GB
Key Board	-	Standard Windows Keyboard
Mouse	-	Two or Three Button Mouse

Monitor - SVGA

S/W System Configuration:-

Operating System : Windows95/98/2000/XP

Application Server : Tomcat5.0/6.X

Front End : HTML, Jsp

Scripts : JavaScript.

Server side Script : Java Server Pages.

Database : MySQL 5.0

Database Connectivity : JDBC