

**Heterogeneous Metric Learning of Categorical Data with Hierarchical Couplings**

**Abstract:**

Learning appropriate metric is critical for effectively capturing complex data characteristics. The metric learning of categorical data with hierarchical coupling relationships and local heterogeneous distributions is very challenging yet rarely explored. This paper proposes a Heterogeneous mEtric Learning with hIerarchical Couplings, HELIC, for this type of categorical data. HELIC captures both low-level value-to-attribute and high-level attribute-to-class hierarchical couplings, and reveals the intrinsic heterogeneities embedded in each level of couplings. Theoretical analyses of the effectiveness and generalization error bound verify that HELIC effectively represents the above complexities. Extensive experiments on 30 data sets with diverse characteristics demonstrate that HELIC-enabled classification significantly enhances the accuracy (up to 40.93%), compared with five state-of-the-art baselines.

**Existing System:**

Most of the existing metric learning methods handle numerical data. Although these methods can learn the distance in numerical data, they cannot handle categorical data directly. While categorical input is involved in work such as they ignore the above-discussed couplings and heterogeneities.

In recent years, several distance metrics or measures have been proposed to capture intra- and inter-attribute couplings in categorical data. For example, the conditional probability and rough membership function capture intra-attribute couplings. The inter-attribute conditional probability and the co-occurrence frequency of highly interdependent attributes measure inter-attribute couplings. A novel categorical data distance measure named coupled object similarity (COS) learns and integrates the intra- and inter-attribute couplings.

**Proposed System:**

First, HELIC captures both low-level value-to-attribute and high-level attribute-to-class couplings to comprehensively reveal the intrinsic and hierarchical characteristics in categorical data. HELIC captures the following interactions: (1) the relationships between the values of an attribute, called intra-attribute couplings, to measure the withinattribute similarities. Such couplings reflect the value interactions within an attribute; (2) the relationships between attributes, called inter-attribute couplings, to measure the between-attribute similarities. These couplings describe the interactions between attribute values conditional on other attributes; and (3) the relationships between attributes and classes, called attribute-class couplings, to measure the attribute-class similarities. These couplings reveal the value distribution w.r.t. each class. Second, HELIC reveals the intrinsic heterogeneities across various types of couplings to identify their different local structures and distributions. Lastly, HELIC learns a heterogeneous metric based on the captured couplings and heterogeneity.