A Weighting Similarity Learning on Categorical Data

Fuyuan Cao[a], Jie Wen[a]

[a]*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China*

Abstract

Attribute independence has been taken as a major assumption in the limited research that has been conducted on similarity analysis for categorical data. However, in real-world data sources,attribute are more or less associated with each other in terms of certain coupling relationships.This paper proposes a weighting distance learning approach that generates a coupled attribute similarity measure for nominal objects with attribute couplings to capture a global picture of attribute similarity.It involves the frequency-based intra-coupled similarity within an attribute and the inter-coupled similarity upon value co-occurrences between attribute as well as their integration on the object level.Substantial experiments on extensive UCI data sets verify the theoretical conclusions.The experimental results show that the similarity measure proposed in this paper has a good effect on clustering data clustering.

Keywords: Clustering, Coupled attribute similarity, Weighting similarity learning

1. Introduction

Similarity analysis has been a problem of great practical importance in several domains for decades, not least in recent work, including behavior analysis, document analysis, and image analysis. A typical aspect of these applications

*Corresponding author

Email addresses: cfy@sxu.edu.cn (Fuyuan Cao), 1967688145@qq.com (Jie Wen)

is clustering. The similarity between clusters is often built on top of the similarity between data objects. The similarity between attribute values assesses the relationship between two data objects and even between two clusters. The more two objects or clusters resemble each other, the lager is the similarity. The other similarity between attributes can also be converted into the difference of similarities between pairwise attribute values. Therefore, the similarity between attribute values plays a fundamental role in similarity analysis.

Compared with the intensive study on the similarity between two numerical variables, such as Euclidean and Minkowski distance, the similarity for categorical data has received much less attention. Only limited efforts have been made, including SMS, which uses 0s and 1s to distinguish the similarity between distinct and identical categorical values, occurrence frequency (OF) and information-theoretical similarity (Lin), to discuss the similarity between nominal values. The challenge is that these methods are too rough to precisely characterize the similarity between categorical attribute values, and only deliver a local picture of the similarity. In addition, none of them provides a comprehensive similarity between categorical attributes by combining relevant aspects.A real database application example is described in Table 1.

Table 1: INSTANCE OF THE MOVIE DATABASE

Movie	*Director*	*Actor*	*Genre*	*Class*
Godfaher II	Scorsese	De Niro	Crime	L1
Good Fellas	Coppola	De Niro	Crime	L1
Vertigo	Hitchcock	Stewart	Thriller	L2
N by NW	Hitchcock	Grant	Thriller	L2
Bishop's Wife	Koster	Grant	Comedy	L2
Harvey	Koster	Stewart	Comedy	L2

As shown in Table 1, six movie objects are divided into two classes with three nominal attributes. The SMS measure between directors Scorsese and Coppola is 0, but Scorsese and Coppola are very similar. Can Wang has put forward an effective method to improve the shortcomings of the above algorithms.

1.1.