Efficient Nearest Neighbor Classification on Categorical Data

Hari Seetha a, V V R Srividya a

aSchool of Computing Science and Engineering, VIT University, Vellore 632014, India, Contact:hariseetha@gmail.com

The similarity measures for continuous data have been well explored when compared to the similarity measures for categorical data. The different attributes in a dataset have different nature and till now no attempt was made to perform classification by applying different similarity measures for different attributes of a dataset. So, in the present paper k-Nearest Neighbor classification is performed using a single similarity measure across all the attributes of each dataset as well as using different similarity measures for different attributes called as hybrid similarity measure. The experimental results on benchmark datasets have shown that classification using hybrid similarity measure outperformed conventional classification.

Keywords : Categorical Data, Hybrid Similarity Measure, k-Nearest Neighbor.

1. INTRODUCTION

The Nearest Neighbor classification depends on the similarity between the objects. Measuring similarity for two data sets is based on several feature variables. This knowledge about similarity is necessary for data mining, pattern recognition, machine intelligence etc. Measuring similarity for categorical data is a challenging problem because they do not have structures. Hence there exists few similarity measures for categorical data. Overlap measure was one of the simplest similarity measure which is defined as \( d(x_i, y_i) = 1 \) if \( x_i = y_i \) else \( d(x_i, y_i) = 0 \)[1]. It simply counts the number of attributes that match in the two data instances. Later, Value Difference Metric (VDM) is used to measure the distance between two categorical values, with respect to class column(supervised learning) [2]. It is defined as:

\[
d(x_i, y_i) = w(x_i) \sum_{c \in C} \left( p(c|x_i) - p(c|y_i) \right)^2
\]

where, \( C \) is the set of all classes labels, \( p(c|x_i) \) is the conditional probability of class \( c \) given \( x \), and \( w(x_i) = \sqrt{\sum_{c \in C} p(c|x_i)^2} \) which attempts to give higher weight to an attribute value that is useful in class discrimination.VDM takes the advantage of the class information, so it is a supervised method. VDM is modified and proposed as Modified Value Distance (MVDM) metric[2], Esposito[3,4] modified traditional hamming distance and various similarity measures e.g., overlap measure, Jaccard(S-coefficient) similarity measure, Sokal-Michener(M-coefficient) similarity measure, Grower-Legendre similarity measure etc., were suggested to get the similarity or dissimilarity coefficient between two categorical data objects. Goodall proposed another statistical approach, in which less frequent attributes have greater contribution to overall similarity than frequent attribute values [5,6]. The Goodall1 measure is the same as Goodall’s measure on a per-attribute basis. However, instead of combining the similarities by taking into account dependencies between attributes, the Goodall1 measure takes the average of the per attribute similarities. Shyam Boriah et al.[6] proposed Goodall3 and Goodall4 which are the other variants of Goodall’s measure. Shoji Hirano et al.[7] adopted the hamming distance that counts the number of attributes for which two objects have different attribute values, in order to measure similarity for categorical attributes,

\[
d_H (x_i, x_j) = \frac{1}{p_H} \sum_{k=1}^{p_H} \delta (x_i^k, x_j^k)
\]
attributes, they showed an improved performance on combining with other similarity measures using hybrid similarity methods.

- The effect of similarity measure on various characteristics of categorical dataset needs to be further explored.

REFERENCES


Hari Seetha is currently an Assistant Professor (Selection Grade) in the School of Computing Science and Engineering at VIT University, Vellore, India and is presently pursuing her Ph.D. She obtained her Master’s Degree in Engineering Physics with specialization in electronics from National Institute of Technology (formerly R. E. C.) Warangal and M.Phil Degree
Nearest Neighbor Classification on Categorical Data

from VIT University, India. She has research interests in the fields of Pattern Recognition, Data Mining, Text Mining and Machine Learning. She has about 10 years of research experience and 19 years of teaching experience. She has published a few research papers in national and international journals and conferences. She had been a co-investigator to a major research project sponsored by the Department of Science and Technology, Government of India.

Srividya V V R completed her Post Graduation in Computer Science at VIT University and is currently working at CTS, Chennai.