International Journal of Information Processing, 6(4), 75-81, 2012
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

Efficient Nearest Neighbor Classification on Categorical Data

Hari Seetha [a], V V R Srividya [a]

[a]School 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:

$$d(x_i, y_i) = w(x_i) \sum_{c \in C} \left(p(c|x_i) - p(c|y_i) \right)^2 \quad (1)$$

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,

$$d_H(x_i, x_j) = \frac{1}{p_H} \sum_{k=1}^{p_d} \delta\left(x_i^k, x_j^k\right) \quad (2)$$

Table 5
CA% vs k using Hybrid3 Similarity Measures

Dataset	Combinations	k=3	k=5	k=10	k=20	k=30	k=50	Max	Min
Car Evaluation	Overlap, OF,Goodall3, Goodall4, Eskin, IOF	88	83	78	90	76	61	90	61
	Overlap, Goodall3, Goodall4, Eskin, IOF, OF	85	93	85	88	85	76	93	76
	Eskin, Overlap, IOF, OF, Goodall3, Goodall4	85	90	93	93	71	56	93	56
	Eskin, Goodall3, Goodall4, Overlap, IOF, OF	90	83	90	76	80	73	90	73
	Eskin, Goodall4, Overlap, IOF, OF, Goodall3	66	56	59	66	51	51	66	51
	IOF, Overlap, Eskin, OF, Goodall3, Goodall4	85	90	93	90	85	71	93	71
	OF, IOF, Goodall3, Goodall4, Overlap, Eskin	93	95	93	88	93	73	95	73
	Goodall3, Overlap, Eskin, IOF, OF, Goodall4	83	90	76	85	59	63	90	59
	Goodall3, IOF, OF, Goodall4, Overlap, Eskin	93	95	93	88	93	73	95	73
	Goodall3, Eskin, IOF, OF, Goodall4, Overlap	56	63	68	68	56	51	68	51
	Goodall4, IOF, OF, Goodall3, Overlap, Eskin	95	93	93	93	93	71	95	71
Chess	Overlap, Goodall3, Goodall4, Eskin, IOF, OF	70	72	64	69	60	56	72	56
	Overlap, Eskin, IOF, OF, Goodall3, Goodall4	48	47	43	53	41	40	53	41
	Overlap, IOF, OF, Goodall3, Goodall4, Eskin	56	53	45	59	43	40	59	40
	Eskin, IOF, OF, Goodall3, Goodall4, Overlap	68	68	61	72	55	49	72	49
	Eskin, Goodall3, Goodall4, Overlap, IOF, OF	70	70	67	72	62	58	72	58
	IOF, Overlap, Eskin, OF, Goodall3, Goodall4	65	64	57	71	55	54	71	54
	Eskin, Goodall4, Overlap, IOF, OF, Goodall3	67	68	63	70	59	57	70	57
	IOF, Eskin, OF, Goodall3, Goodall4, Overlap	73	72	67	75	63	60	73	63
	IOF, OF, Goodall3, Goodall4, Overlap, Eskin	72	70	64	76	59	57	76	57
	IOF, Goodall3, Goodall4, Overlap, Eskin, OF	77	78	73	77	71	67	78	67
	IOF, Goodall4, Overlap, Eskin, OF, Goodall3	70	69	65	74	64	61	74	61
	Goodall4, Overlap, Eskin, IOF, OF, Goodall3	78	76	73	77	68	63	78	63
	Goodall4, Goodall3, Overlap, Eskin, IOF, OF	70	70	65	71	61	57	71	57

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

1. Stanfill C and Waltz D. Toward Memory-Based Reasoning. *Communications. ACM*, 9(12): 1213-1228,1986.
2. Cost S and Salzberg S. A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features, *Machine Learning*, 10(1): 57-78, 1993.
3. Esposito F, Malebra D, Tamma V, Bock H H. Classical Resemblance Measures, *Analysis of Symbolic Data. Springer*, pages 139-152, 2002.
4. Aamir Ahmad, Lipika Dey, A Method to Compute Distance between Two Categorical Values of same Attribute in Unsupervised Learning for Categorical Data Set, *Pattern Recognition Letters,* 28(1): 110-118, 2007.
5. Goodall D W. A new Similarity Index Based on Probability. *Biometrics*, 22(4): 882-907, 1966.
6. Shyam Boriah, Varun Chandola and Vipin Kumar, Similarity Measures for ategorical Data: A Comparative Evaluation, *In Proceedings of 2008 SIAM Data Mining Conference, April 2008, Atlanta, GA*, pages 243-254, 2008.
7. Shoji Hirano, Xiaoguang Sun, Shusaku Tsumoto. On Similarity Measures for Cluster Analysis in Clinical Laboratory Examination Databases. *In Compsac, 26th Annual International Computer Software and Applications Conference*, page 1170, 2002.
8. UCI Machine Learning Repository. *http://archive.ics.uci.edu/ml/datasets.html*

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

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.