A Hybrid Approach to Classification of Categorical Data Based on Information-Theoretic Context Selection

Madhavi Alamuri, Bapi Raju Surampudi and Atul Negi

Abstract Clustering or classification of data described by categorical attributes is a challenging task in data mining. This is because it is difficult to define a measure between pairs of values of a categorical attributes. The difficulty arises due to lack of ordering information between various pairs of categorical attributes. In this paper we introduce a Hybrid Approach which combines set based context selection with distance computation using KL divergence method. In the literature context based approaches have been introduced recently. Current approaches look at categorical attributes individually, however our approach proposes a novel scheme inspired from information theory. We consider the interdependence redundancy measure to select the significant attributes for context selection. The proposed approach gives encouraging results for low dimensional benchmark UCI datasets with k-nearest neighbor classifier based on the proposed measure. On these datasets the proposed measure performed well in comparison to other distance measures while using various classifiers such as SVM, Naive Bayes and C4.5.

Keywords Categorical data · Similarity · Context · Classification

1 Introduction

Similarity or distance between two objects plays a significant role in many data mining and machine learning tasks like classification, clustering and outlier detection. In general distance computation is a built-in step for these learning algorithms and different distance measures can be conveniently used. However the effectiveness of

M. Alamuri (✉) · B.R. Surampudi · A. Negi
School of Computer and Information Sciences, University of Hyderabad,
Hyderabad, India
e-mail: madhavi_alamuri@yahoo.com

B.R. Surampudi
Cognitive Science Lab, International Institute of Information Technology,
Hyderabad, India

V. Ravi et al. (eds.), *Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO - 2015)*, Advances in Intelligent Systems and Computing 415, DOI 10.1007/978-3-319-27212-2_22

the proposed distance measure/metric usually has significant influence on the tasks like classification and clustering.

For numerical data sets, the distance computation is well understood and most commonly used measures such as minkowski distance, mahalanobis distance can be applied. However, measuring similarity or distance is not straight forward for categorical data sets as there is no explicit ordering of categorical values and it is very difficult to determine how much one symbol differs from another. By categorical data we mean the values that are nominal e.g. $color = \{black, blue, green\}$ or ordinal e.g. $size = \{small, large, verylarge\}$.

In this paper we study the similarity/distance measures for categorical data objects and propose a hybrid approach, based on Set based Context Selection (SBCS) and distance computation on the context using KL divergence method. In our Hybrid Approach (HA) context for each attribute refers to the subset of attributes which gives some contextual interpretation over the attribute set.

The proposed approach can well quantify the distance in supervised learning environment. In this paper we also focus on a data driven methods for selecting a good context for a given attribute. We provide information theoretic approaches for context selection of each and every attribute. The underlying assumption in our approach is that if the similarity function has high value for the given context, then the objects represented by the given context description are similar. Recently, increasing attention is being paid to find the grouping structure/classification of categorical data.

The rest of the paper is organized as follows: In Sect. 2 we discuss the state of the art in categorical similarity/dissimilarity measures. In Sect. 3 we present the theoretical details of Hybrid Approach. In Sect. 4 we present the technical details and in Sect. 5 we present the results of set of experiments on low dimensional UCI benchmark datasets.

2 Background

According to Michalski [16], the conventional measures of similarity are "*Context-free*" i.e. the distance between any two data objects X and Y is a function of these points only, and does not depend on the relationship of these points to other data points.

$$Similarity(X, Y) = f(X, Y)$$

Context-free similarity measures may be inadequate in some clustering/classification problems. Hence recent approaches for finding similarity measures include "*Context-sensitive*" methods, where

$$Similarity(X, Y) = f(X, Y, Context)$$

Here the similarity between X and Y depends not only on X and Y, but also on the relationship of X and Y to other data points, represented by Context.

In a user driven approach context refers to the subset of attributes of interest and is application dependent.

Pearson's [17] Chi-square statistic was often used in the late 1800s to test the independence between categorical variables in a contingency table. Sneath and Sokal [19] were among the first to put all the similarity measures together. The conventional methods of similarity measures used to binarize the attribute values where bit 1 indicates the presence and bit 0 indicates absence of a possible attribute value. After obtaining binary vectors, binary similarity measures are used to apply on them. The major drawback is transforming objects into a binary vector may leave many minute insights in to the dataset. The most Commonly used similarity measures for categorical data is overlap, where similar values are assigned a distance of 1, and dissimilar values are assigned a distance of 0. The drawback with this measure is, it does not distinguish the different values taken by an attribute.

In general the similarity measures for categorical data are categorized based on how they utilize the context of the given attributes. There are several supervised and unsupervised measures from the literature are existing to find the similarity between categorical feature values.

- The supervised similarity measures do consider the class attribute information. The supervised measures are further divided into learning, non-learning approaches. The learning approaches can be IVDM [21], WVDM [21], Learned Dissimilarity measure [22]. The non-learning approaches can be VDM [20], MVDM [7].
- The unsupervised similarity measures do not consider the class attribute information. These measures are further classified into Probabilistic, Information theoretic, and Frequency based approaches.
  - Probabilistic approaches: Goodall [9], Smirnov [18], Anderberg [2].
  - Information theoretic approaches: Lin [15], Lin1 [4], Burnaby [6].
  - Frequency based approaches: OF [11], IOF [11].

Boriah et al. [4] classified the similarity measures based on the storage mechanism of similarity values in the similarity matrix, parameters used to propose the measure and based on the weight of the frequency of the attribute values.

Jierui Xie [22] proposed a learning algorithm, which learns a dissiimilarity measure by mapping each categorical value into random numbers. This learning algorithm is guided by the classification error for effective classification of the data points.

The taxonomy of various distance/similarity measures for categorical data is explored in [1].

Apart from these categories of similarity measures, recent attention has been paid to context-based approaches. Dino et al. [10] present a context based approach to compute distance between pair of values of a categorical data. They proposed a distance learning framework for context selection and validated in a hierarchical clustering algorithm.

Zeinab et al. [12] proposed a novel approach for distance learning based on the context information. This method is also used to compute dissimilarity between probability distributions. Both these approaches use the context information for dissimilarity computation.

We extend the current context based approaches by introducing a Hybrid Approach which utilizes information-theoretic measures. The proposed approach is explored in Sect. 3.

3 Proposed Hybrid Approach

In this section we present a Hybrid Approach for computing distance between any pair of values of a categorical attribute. We also introduce "Set Based Context Selection Algorithm" (SBCS) for the effective selection of the context followed by calculating the distance by using KL divergence [13, 14] method. Our Hybrid Approach is formulated based on the following two steps.

1. Set Based Context Selection: This method selects a subset of correlated features based on a given attribute. i.e., selection of a meta attribute set of a given attribute which is relevant in terms of information theoretic measure is calculated in this step.
2. Distance Computation: Computation of the distance measure between pair of values of an attribute using the meta attributes set defined in the previous step. KL divergence method is applied on the context to measure the difference between the probability distributions.

The essential premise in formulating this algorithm is with an open minded, fairness in the importance of the attributes, it excludes weightage or bias towards a certain set of attributes, unless it is explicitly defined in the context.

The notations used in this algorithm are as follows: Consider the set $F = \{A_1, A_2, \ldots A_m\}$ of m categorical attributes and let the set of instances set $D = \{X_1, X_2 \ldots X_n\}$. We denote by a lower case letter $a_i \epsilon A_i$, a specific value of an attribute A_i.

3.1 Set Based Context Selection

The selection of a good context is not trivial, when data are high dimensional. In the SBCS, we use mutual information normalized with joint entropy to get the context for each and every attribute.

The Set Based Context Selection Algorithm which we propose considers a score for each feature independently of others. A useful and relevant set of features may not only be individually relevant but also may not be redundant with respect to each other. The selecting criterion of a context is based on the relevance index which quantifies whether a particular feature can be included in a context set or not.

In the following sub section we introduce some basic concepts of information theoretic measures followed by how they are utilized to tackle the problem of context selection.

3.2 Entropy and Mutual Information

The *entropy* of a random variable [8], is the fundamental unit of information which quantifies the amount of uncertainty present in the distribution of the random variable.

The entropy of a random variable A_i is defined as,

$$H(A_i) = -\sum_{k \epsilon A_i} p(a_k^i) log_2 p(a_k^i) \tag{1}$$

where $p(a_k^i)$ is the probability of value of a_k of attribute A_i.

The entropy of a random variable can be conditioned on other variables. The conditional entropy of A_i given A_j is,

$$H(A_i|A_j) = -\sum_{k \epsilon A_j} p(a_k^j) \sum_{l \epsilon A_i} p(a_l^i|a_k^j) log_2 p(a_l^i|a_k^j) \tag{2}$$

where $p(a_l^i|a_k^j)$ is the probability that $A_i = a_l$ after observing the value $A_j = a_k$. This can be interpreted as the amount of uncertainty present in A_i after observing the variable A_j.

The amount of Information shared between A_i and A_j, which is also called as mutual information is defined by,

$$I(A_i; A_j) = H(A_i) - H(A_i|A_j) \tag{3}$$

This is the difference between two entropies which can be interpreted as the amount of uncertainty in A_i which is removed by knowing A_j.

The mutual information between two attributes also measures the average reduction in uncertainty with another attribute. A smaller value of mutual information indicates lesser dependence and a larger value of mutual information indicates greater dependence.

The main drawback of using this measure is that the mutual information value increases with the number of distinct values that can be chosen by each attribute. To overcome this problem Au et al. [3] proposed interdependence redundancy measure where mutual information is normalized with joint entropy, which is defined as,

$$IDR(A_i, A_j) = \frac{I(A_i; A_j)}{H(A_i, A_j)} \tag{4}$$

where the joint entropy $H(A_i, A_j)$ is calculated as,

$$H(A_i, A_j) = -\sum_{k\epsilon A_i}\sum_{l\epsilon A_j} p(a_k^i, a_l^j) log_2 p(a_k^i, a_l^j) \tag{5}$$

According to [3] the interdependence measure evaluates the degree of dependency between two attributes. Unlike mutual information, where the number of possible values which an attribute can take effect, the interdependency measure has no effect on the number of distinct values taken by an attribute. Hence IDR measure is considered as more ideal index to rank the attributes in terms of dependency. $IDR(A_i, A_j) = 1$ means that the attributes A_i and A_j are dependent on each other while $IDR(A_i, A_j) = 0$ indicates that the attributes are statistically independent. When the value of IDR lies between 0 and 1 the attributes are partially independent. By using this IDR measure, we can maintain a $m \times m$ matrix IDR to store the dependency degree of pair of attributes.

For each and every attribute A_i we find all the attributes that have interdependency with it and store them in a context set. In order to not to unnecessarily increase the size of the context set we introduce a threshold t to include the significant attributes in the context set.

$$context(A_i) = \left\{A_k | IDR(A_i; A_k > t, A_i, A_k \epsilon F)\right\} \tag{6}$$

It is being conjectured that for lower values of threshold t we may obtain higher values of classification accuracy. However for larger values of the threshold the classification accuracy may drop.

The relationship between these quantities is explored in [5] and can be observed from the Fig. 1.

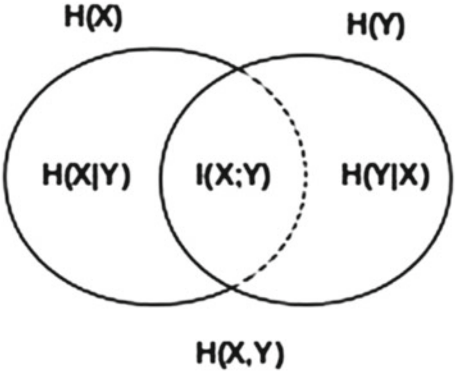

Fig. 1 Relationship of various information-theoretic measures

To get the context of a given attribute we make use of these information-theoretic measures by adding/removing the relevant features in the context set.

3.3 Distance Computation

The distance between pair of values (x_j, y_j) of a categorical attribute A_j is formulated using KL divergence method as,

$$d_{A_j}(x_j, y_j) = \sum_{A_i \epsilon context(A_j)} \sum_{v_i \epsilon A_i} p(v_i|x_j) log \frac{p(v_i|x_j)}{p(v_i|y_j)} + p(v_i|y_j) log \frac{p(v_i|y_j)}{p(v_i|x_j)} \quad (7)$$

The distance defined above depends on the meta attribute set associated to each attribute, where meta attribute set is derived from the SBCS algorithm. The dissimilarity between two probability distributions using KL divergence is symmetric, and hence the distance between pair of values of an attribute is also symmetric.

Let X and Y be two instances of the dataset then the above calculated pairwise distance between attribute values is embedded in the total distance as,

$$D(X, Y) = \sum_{j=1}^{m} d_{A_j}(x_j, y_j) \quad (8)$$

4 Hybrid Approach Implementation

In this section we introduce the algorithmic details for the implementation of (i) Set Based Context Selection and (ii) Distance Computation of Hybrid Approach.

In Algorithm 1 we propose interdependence redundancy based context selection for each attribute in the feature set. Initially the selected set S is empty and the unselected set US is the set of all features from the feature set. At line 6 this algorithm selects an attribute to be included in the context set based on the threshold t and at line 8, it deselects the attribute from the unselected set US. When the context is chosen for a given attribute Distance Computation function computes the distance matrix between each pair of values of the attribute.

In Algorithm 2 distance measure is computed between pair of values of $A_i \epsilon F$ using context set derived from the first step of the Hybrid Approach.

The total distance between two objects is then calculated by using Eq. 8 defined in Sect. 3.3.

Algorithm 1 Hybrid Algorithm

```
procedure HYBRID APPROACH($D, F$)
    selectedSet S = {};
    unselectedSet US = {$A_1, A_2...A_m$};
    For each pair of attributes $(A_i, A_j), (i, j)\epsilon\{1, 2, ..., F\})$ calculate $IDR(A_i, A_j)$
    according to Eq. 4
    for all $A_i \epsilon F$ do
        a. Find the feature $A_k$ from the US such that $IDR(i, k) > t$
        b. insert($A_k$,S)
        c. remove($A_k$,US)
        d. contextSet($A_i$) = S
    end for
    $DistanceMatrix_{A_i} = calculateDistance(A_i, contextSet(A_i))$;
    return $DistanceMatrix_{A_i}$
end procedure
```

Algorithm 2 Distance Computation

```
procedure CALCULATEDISTANCE($A_j$, $Context(A_j)$)
    for all $x_j, y_j \epsilon A_j$ do
        if $x_j \neq y_j$ then
            $d_{A_j}(x_j, y_j) = \sum_{A_i \epsilon context(A_j)} \sum_{v_i \epsilon A_i} p(v_i|x_j) log \frac{p(v_i|x_j)}{p(v_i|y_j)} + p(v_i|y_j) log \frac{p(v_i|y_j)}{p(v_i|x_j)}$
        else
            $d_{A_j}(x_j, y_j) = 0$
    end for
    return $d_{A_j}$
end procedure
```

5 Experimental Results

To evaluate the proposed Hybrid Approach, we compare our approach with other base-line similarity measures explored in Sect. 2 and with the other classifiers. We present results on 5 benchmark categorical datasets, which are taken from the UCI machine learning repository (Table 1).

5.1 Results Description

We compare our Hybrid Approach with the 5 similarity measures Overlap, Lin, Gambaryan, OF, IOF described in Sect. 2. We evaluate the classification accuracy of the nearest neighbor classifier (k = 7) with five fold cross validation. The last row of the Tables 2 and 3 gives the average performance over all the datasets. In summary

Table 1 Dataset description

Dataset	Size	Dimension	Attributes and symbols	No.of classes
Mushroom	8124	22	Various sizes from 2 to 12, e.g. $cap - surface = \{fibrous, grooves, scaly, smooth\}$	2
Tic-tac-toe	958	9	Each attribute takes on $\{x, o, b\}$	2
Balance scale	625	4	Each attribute takes on $\{1, 2, 3, 4, 5\}$	3
Car evaluation	1728	6	Each attribute takes different values e.g. $buying = \{vhigh, high, med, low\}$	4
Hayes-Roth	160	5	Each attribute takes on different values e.g. $hobby = \{1, 2, 3\}$	3

Table 2 Performance comparison with various similarity measures with knn (k = 7)

Dataset	Overlap	Lin	Gambaryan	OF	IOF	HA
Mushroom	**100**	98.75	53.00	98.9	99.95	97.23
Tic-tac-toe	82.35	**97.30**	80.45	72.49	95.13	92.34
Balance Scale	72.31	72.21	72.32	73.59	72.34	**85.12**
car Evaluation	88.34	**93.10**	83.20	92.43	87.13	90.46
Hayes-Roth	68.50	71.00	67.52	60.00	65.50	**83.55**
Average	82.3	86.472	71.298	79.482	84.01	**89.74**

Table 3 Performance comparison with various classifiers

Dataset	SVM	NB	C4.5	HA
Mushroom	**100**	96.5	**100**	98
Tic-tac-toe	77	75.34	84.12	**92**
Balance scale	95.5	**96.3**	73	88.21
Car evaluation	88.2	92.1	**96.51**	95.35
Hayes-Roth	64	68.5	71	**80**
Average	84.94	85.748	84.926	**90.712**

the proposed Hybrid Approach achieves best rank in two datasets namely Balance Scale and Hayes Roth and stands best on average classifier accuracy.

We also compare our Hybrid Approach with the algorithms implemented in Weka 3.6.10 including SVM, C4.5 and Naive Bayes (NB). Our method uses the set based context selection with KL divergence as a distance measure whereas the other methods use the Euclidean distance with simple matching between categorical objects. As shown in Table 3, the Hybrid Approach performs better in Tic-Tac-Toe and Hayes-Roth and stands best in average performance with 5 datasets.

6 Conclusions

In this paper we propose a hybrid approach to measure similarity between categorical attribute values. This algorithm uses *Set Based Context Selection* method inspired from information-theoretic measures. We tested our approach on five benchmark datasets from UCI machine learning repository. The proposed approach gives promising results for low-dimensional datasets.

The proposed approach gives superior results for datasets with dimensionality approximately below 10. Computation of context selection is very expensive for high dimensional datasets. This is a limitation of the proposed approach. We are exploring generalization of dimensionality reduction techniques for categorical attributes so that the proposed approach can be combined with these methods in future. We also investigate the impact of the threshold parameter t on the proposed distance measure for large datasets in future.

References

1. Alamuri, M., Surampudi, B.R., Negi, A.: A survey of distance/similarity measures for categorical data. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 1907–1914. IEEE (2014)
2. Anderberg, M.R.: Cluster analysis for applications. Technical report, DTIC Document (1973)
3. Au, W.H., Chan, K.C., Wong, A.K., Wang, Y.: Attribute clustering for grouping, selection, and classification of gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB) **2**(2), 83–101 (2005)
4. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: a comparative evaluation. In: Proceedings SIAM International Conference on Data Mining. SIAM, Atlanta (2008)
5. Brown, G., Pocock, A., Zhao, M.J., Luján, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J. Mach. Learn. Res. **13**(1), 27–66 (2012)
6. Burnaby, T.P.: On a method for character weighting a similarity coefficient, employing the concept of information. J. Int. Assoc. Math. Geol. **2**(1), 25–38 (1970)
7. Cost, S., Salzberg, S.: A weighted nearest neighbor algorithm for learning with symbolic features. Mach. Learn. **10**(1), 57–78 (1993)
8. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, New York (1991)
9. Goodall, D.W.: A new similarity index based on probability. Biometrics **22**(4), 882–907 (1966)
10. Ienco, D., Pensa, R.G., Meo, R.: From context to distance: learning dissimilarity for categorical data clustering. ACM Trans. Knowl. Discov. Data (TKDD) **6**(1), 1 (2012)
11. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. J. Documentation **28**(1), 11–21 (1972)

12. Khorshidpour, Z., Hashemi, S., Hamzeh, A.: Distance learning for categorical attribute based on context information. In: 2010 2nd International Conference on Software Technology and Engineering (ICSTE), vol. 2, pp. V2–296. IEEE (2010)
13. Kullback, S.: Information Theory and Statistics. John Wiley & Sons, New York (1959)
14. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
15. Lin, D.: An information-theoretic definition of similarity. Proc. ICML **98**, 296–304 (1998)
16. Michalski, R.S.: Knowledge acquisition through conceptual clustering: a theoretical framework and algorithm for partitioning data into conjunctive concepts. Int. J. Policy Anal. Inf. Syst. **4**(3), 219–244 (1980)
17. Pearson, K.: On the general theory of multiple contingency with special reference to partial contingency. Biometrika **11**(3), 145–158 (1916)
18. Smirnov, E.S.: On exact methods in systematics. Syst. Biol. **17**(1), 1–13 (1968)
19. Sneath, P.H.A., Sokal, R.: Numerical taxonomy. The principles and practice of numerical classification. Freeman, San Francisco (1973)
20. Stanfill, C., Waltz, D.: Toward memory-based reasoning. Commun. ACM **29**(12), 1213–1228 (1986)
21. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. J. Artif. Intell. Res. **6**, 1–34 (1997)
22. Xie, J., Szymanski, B., Zaki, M.J.: Learning dissimilarities for categorical symbols. J. Mach. Learn. Res. Proc. Track **10**, 97–106 (2010)