International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Clustering Approach using Hierarchical Coupling Learning for Categorical Data

Nilam Patil and Prof. S. M. Kamalapur

Department of Computer Engineering K. K. Wagh Institute Of Engineering Education & Research, Nashik

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

A bulk data is generated from various sources. Many real world applications generate categorical data with finite unordered feature values. Like numerical data categorical data cannot be directly processed using algebraic operation. Hence many machine learning numerical processing algorithms cannot be directly applied to the categorical dataset. The categorical data is converted in the numerical form and then such numerical machine learning algorithms can be applied. A lot of work has been done in literature for data representation. For good data representation intrinsic data characteristics should be effectively captured. Some technique in literature focuses on low level strong coupling between feature values while other are focusing on high level clusters of feature values. In the proposed system Coupled Unsupervised categorical data Representation (CURE) framework is proposed. It uses hierarchical learning structure. It defines value to value as well as value clusters coupling. Along with the data representation the Coupled Data Embedding(CDE) algorithm is proposed to generate clusters of categorical data using numerical representation. The clusters of numerical data are generated using k-menas algorithm. The results are compared with clustering of categorical data using k-mode algorithm. Categorical data Representation, clustering, Unsupervised Learning, kmeans

Keywords: *Categorical data Representation, clustering, Unsupervised Learning, kmeans*

Introduction

Categorical data is generated in variety of applications. The categorical data with nominal attributes i.e. attribute contains finite set of values are appearing in various real world applications. The numerical manipulations cannot be directly applied on categorical dataset. The various data mining algorithms require numerical representation of data. On numerical data representation various operations can be performed like clustering, classification and regression. It is important to convert the categorical data to the numerical form for further operations.
The good representation of categorical data should preserve the essential data properties. During data conversion the various coupling information among data should be preserved. The coupling is categorized in 2 sections:

1. Low level coupling: In low level coupling, the relationships among various attribute values are identified. The coupled values are generally co-occurred in various data instances. Consider an example: The education feature value “PhD” is generally co-occurred with “professor” or “scientist” value in the occupation attribute. Such various attributes has its intrinsic relationship and form a semantic value clusters.
2. High level coupling: In the high level coupling the value clusters are further coupled with each other. The clustering of all feature values is performed with different granularity.

In the existing work the coupling among values i.e. coupling at low level is not considered. For coupling information supervised dataset is required. In variety of situation the data is unsupervised and there is need to convert such data to numerical form by preserving its intrinsic coupling information.
For unsupervised data, clustering is important functionality to further analyze the information. This is required in variety of domain such as medicines, computer vision, biology, marketing etc. Unlike the statistical methods the clustering is independent of learning process. It does not require any training or any pre-assumptions to describe the underlying data structure.
The proposed method proposes a technique to convert the categorical data to the numerical form by preserving the intrinsic properties of data. Based on the numerical representation the clustering technique is applied to generate clusters of categorical data. For clustering more relevant information from data should

be captured so that more accurate clusters can be generated.
Following section includes the study of related work in the domain of categorical data representation and clustering. Based on the analysis of existing methods the problem statement is proposed in section III. The details of proposed system is given in section IV followed by the result and analysis and conclusion.

Related Work

The most widely used method is encoding method. This method is used for categorical data representation. There are various methods such as One Hot Encoding, Label Encoding, Frequency Encoding, Probability Ratio Encoding, etc.[21].
In the One Hot Encoding[22] the feature value is converted in to 0 1 matrix. The distinct values of attribute are treated as an individual feature. Based on the occurrence of value in the instance the feature value is set to 1 and rest entries are kept 0. This is reversible technique i.e. form numerical representation data can be regenerated. It assumes that all data values are independent.
In IDF encoding[8], each value is represented with logarithm of its inverse frequency. In this technique coupling information is captured with occurrence frequency. This technique does not capture the complex value coupling information. This is efficient technique for generating numerical representation of data.
For textual data conversion some methods like latent semantic indexing (LSI) [24], latent Dirichlet allocation(LDA) [25] are available. But categorical data has different structure than unstructured textual form data. These methods cannot be directly applied to the categorical data.
To find value coupling between data objects, the similarity learning measures are proposed. These measures find object to object similarity matrix. ALGO_DISTANCE [9] technique is used to find object to object coupling information based on the conditional probability. The DIstance Learning for Categorical Attributes (DILCA)[10] similarity measure finds the similarity of feature objects based on the feature selection and feature weighting technique. For feature selection it uses Symmetric Uncertainty. For feature weighting it uses context selection of features. Distance Metric(DM)[11] uses frequency probabilities and attribute-distance for similarity measurement. All these methods are failed to capture the coupling among multiple values in dataset and the relationship among cluster of values.
To overcome the limitation of existing approaches a CURE framework is proposed. This framework focuses on extraction of coupling information. The framework Learns: Value Coupling, Value Clusters, Couplings between Value Clusters and then Object Representation. Based on the coupling information CDE algorithm is proposed and numerical representation of categorical data is generated. It uses Principal component analysis (PCA) technique for removing less discriminative features form dataset. PCA is a linear projection of greatest variance from top eigenvectors of covariance matrix of data. This PCA structure do not preserves the low dimensional embedding of data and pair wise distance between data points.

Problem Formulation

Lot of applications generates categorical data. This data cannot be directly manipulated. There is need to convert this data in numerical form for further manipulation. Various techniques in literature are proposed to generate numerical representation of data. For good data representation intrinsic data characteristics should be effectively captured. Some technique in literature focuses on low level strong coupling between feature values while other are focusing on high level clusters of feature values.
The CURE framework preserves the intrinsic data properties by analyzing value to value coupling, value clusters and value cluster coupling. The CDE algorithm follows CURE framework but this algorithm uses principal component analysis technique for removing less discriminative features. These PCA techniques do not preserve the pair wise distance between data points. To preserve intrinsic property of data, new technique for dimensionality reduction is required that preserves the pair wise distance between data points.

Proposed System

A. System Architecture

Following Figure 3 shows the architecture of the system. The categorical dataset is input to the system. The coupled data embedding (CDE) and Local Linear Embedding algorithms is applied to the system to generate numerical representation of data. The CDE algorithm follows the coupled unsupervised categorical data representation (CURE) framework. This framework preserves the intrinsic data properties while converting it to the numerical form. After numerical conversion Kmeans clustering is applied on dataset and F-score is evaluated. Following figure shows the architecture of the system.

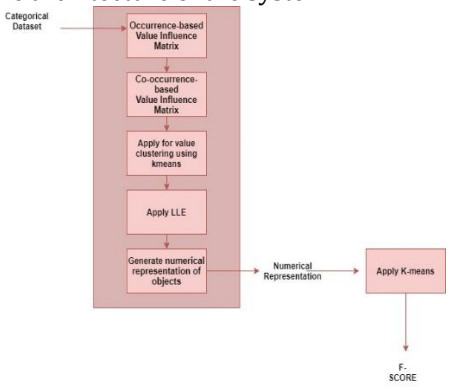

Fig 1 System Architecture

B. Preliminaries:

1. Joint Probability :

The joint probability of two values u and v can be calculated as:

$$P(v_i, v_j) = \frac{|\{ v_x^{fi} = v_i \cap v_x^{fj} = v_j \ , \ x \in X \}|}{N} \quad (1)$$

F is is feature in dataset and vxfi and vxfj are two values of instance x of features i and j.

N : total number of instances.

2. Normalized mutual information:

Normalized mutual information between two features a and b is given as :

$P(Vi, Vj)$

Ψ (

$$f_a, f_b) = \frac{2 \sum_{v_i \in v_{fa}} \sum_{v_j \in v_{fb}} p(v_i, v_j) \log \frac{}{P(V_i)P(V_i)}}{h(f_a) + h(f_b)} \quad (2)$$

Where, p(vi) is the occurrence frequency of vi

$$h(f_a) = -\sum_{vi \in Vfa} P(v_i) \log(p(v_i)) \quad (3)$$

3. Occurrence-based Value Influence Matrix

The co-occurrence-based value influence matrix Mc is defined as:

$$M_o = \begin{bmatrix} \emptyset_o(v_1, v_1) & \cdots & \emptyset_o(v_1, v_L) \\ \vdots & \ddots & \vdots \\ \emptyset_o(v_L, v_1) & \cdots & \emptyset_o(v_1, v_L) \end{bmatrix}$$

(4)

Where ,

$\emptyset_o(v_i, v_j) = \psi(f_i, f_j) \times \frac{P(v_j)}{P(v_i)}$ is the coupling function.

4. Co-occurrence-based Value Influence Matrix

The co-occurrence-based value influence matrix Mc is defined as :

$$M_c = \begin{bmatrix} \emptyset_c(v_1, v_1) & \cdots & \emptyset_c(v_1, v_L) \\ \vdots & \ddots & \vdots \\ \emptyset_c(v_L, v_1) & \cdots & \emptyset_c(v_1, v_L) \end{bmatrix}$$

(5)

Where ,

$\emptyset_c(v_i, v_j) = \frac{P(v_j, v_j)}{P(v_i)}$ is the coupling function.

C. System Working

The system uses coupled unsupervised categorical data representation (CURE) framework for converting categorical data to numerical format. In this framework initially value coupling is performed. Then it focuses on value clusters. After creating value clusters it finds coupling between value clusters. After analysis numerical representation is generated. The coupled data embedding (CDE) algorithm is proposed. This algorithm follows the CURE framework and generates numerical representation of categorical data.

As per the CURE framework, CDE initially focuses on value coupling feature of dataset. For value coupling, occurrence-based Value Influence Matrix and Co-occurrencebased Value Influence Matrix matrices are generated. After finding the value coupling the value clusters are performed. For value cluster generation Kmeans algorithm is used. Initially the cluster count i.e. value of k is set to 2. Gradually the value increases as per the proportion factor.

After creating value clusters it finds value to value cluster affiliation. The clusters generated using Kmeans are again refined. The clusters with less discriminative information are deleted from cluster list. Then from the filtered clusters list CDE algorithm finds the coupling between value clusters. The attributes with less information are removed using nonlinear dimensionality reduction (NLDR) technique: Local Linear Embedding (LLE). Uisng this technique the dimensions of dataset are reduced and attributes which provide more discriminative information are preserved.

In LLE, initially nearest neighbor of each point are identified. Then it finds the set of weights of each point that describes the point as a linear combination of its neighbors. Then it uses eigen vector based optimization and find lower dimensional representation of points.

The generated lower dimensional matrix is the representation of categorical data in numerical form. A clustering using Kmeans algorithm is applied on this dataset and Fscore is evaluated. In the following section detailed steps of algorithm CDE and LEE are given.

D. Algorithm:

Algorithm 1: CDE

Input :

D - data set,

D: reduced dimension count A: proportion factor

Output:

Y:the numerical representation of objects **Processing:**

1. Generate Mo and Mc Matrix using preliminary 3 and 4.
2. For matrices Mo and Mc
3. Initialize cluster count = 2
4. Initialize CI = EMPTY, k = EMPTY
5. Do
6. Update CI using kmeans(M, k)
7. Store the clusters with one value in Cs
8. Remove the clusters with one value from CI
9. Update Value of K
10. While length(Cs)/k >=A
11. Y: Apply LLE(CI,d) Algorithm
12. Return Y

Algorithm 2: LLE algorithm:

Input :

CI: Dataset with N instances and D dimensions d: Reduced dimension Count **Output :** Y: reduced dimension Set **Processing:**

1. Find neighbors of each data point in CI

2. For i=1 to N
3. Create a matrix Z containing all neighbors of CIi
4. subtract CIi from Z
5. local covariance C=Z'*Z
6. Find w by solving C*w = 1
7. if j is not a neighbor of i
8. set Wij=0
9. else
10. set w/sum(w);
11. create sparse matrix M = (I-W)'*(I-W)
12. find bottom d+1 eigenvectors of M
13. set the qth ROW of Y to be the q+1 smallest eigenvector
14. remove bottom eigen vectors with eigen value 0
15. return Y

Result and Discussions

A] Experimental Setup:

Table 1 : Hardware and Software Requirements

Software Specification:	
Language:	Java – jdk1.8
Development Tool:	NetBeans IDE 8.2
Operating System	Windows 7
Hardware Specification:	
RAM	4GB
Processor	I5

B] Datasets:

UCI[11] benchmark data sets datasets are used for system testing.

Following table gives the detailed description of dataset.

Table 2 : Dataset Description

Sr. No.	Dataset	Number of Attribute	Number of instances	Number of attributes
1.	Zoo	17	101	30
2.	Dermatology	33	366	129
3.	Soybeansmall	3	47	58
4.	Hepatitis	19	156	36

B. Performance Measures:

1. F-score : f-score is evaluated after applying the Kmeans clustering algorithm on numerical representation dataset.

2. Evaluation Time: Execution Time for numerical representation of categorical data is evaluated for various datasets.

C] Implementation Status:

From the given dataset Occurrence-based Value Influence Matrix Mo and Co-occurrence-based Value Influence Matrix Mc are generated.

Conclusions

The proposed system presents a Coupled Unsupervised categorical data Representation (CURE) framework based Coupled Data Embedding(CDE) algorithm. To reduce the dimension space nonlinear dimensionality reduction (NLDR) technique: Local Linear Embedding is used. It defines value to value as well as value clusters coupling. Afterwards the system focuses on coupling between value clusters and after analysis numerical representation is generated. The clusters of numerical data are generated using k-means algorithm. The Fscore is calculated using Kmeans clustering result.

References

[1] onglei Jian, Guansong Pang, Longbing Cao, Kai Lu and Hang Gao, "CURE: Flexible Categorical Data Representation by Hierarchical Coupling Learning", in IEEE Transactions on Knowledge and Data Engineering Vol. 31 , Issue: 5 , May 2019,pp:853 - 866

[2] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, Applied multiple regression/correlation analysis for the behavioral sciences. Routledge, 2013.

[3] Y. Bengio, Y. LeCun et al., "Scaling learning algorithms towards ai," Large-scale kernel machines, vol. 34, no. 5, pp. 1–41, 2007

[4] A. Aizawa, "An information-theoretic perspective of tf–idf measures," Information Processing & Management, vol. 39, no. 1, pp. 45–65, 2003.

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, vol. 41, no. 6, p. 391, 1990.

[6] M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," JMLR, vol. 3, no. Jan, pp. 993–1022, 2003.

[7] A. Ahmad and L. Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," Pattern Recognition Letters, vol. 28, no. 1, pp. 110– 118, 2007.

[8] Ienco, R. G. Pensa, and R. Meo, "From context to distance: Learning dissimilarity for categorical data clustering," ACM TKDD, vol. 6, no. 1, p. 1, 2012.

[9] H. Jia, Y.-m. Cheung, and J. Liu, "A new distance metric for unsupervised learning of categorical data," IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 5, pp. 1065–1079, 2016.

[10] C. Wang, X. Dong, F. Zhou, L. Cao, and C.-H. Chi, "Coupled attribute similarity learning on categorical data," IEEE TNNLS, vol. 26, no. 4, pp. 781–797, 2015.

[11] Datasets: https://archive.ics.uci.edu/ml/datasets.html