International Journal of Research in Engineering, Science and Management
Volume-3, Issue-2, February-2020
www.ijresm.com | ISSN (Online): 2581-5792

A Survey of Clustering Approach using Hierarchical Coupling Learning for Categorical Data

Nilam Patil[1], Snehal Kamalapur[2]
[1]*M.E. Student, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik, India*
[2]*Professor, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik, India*

***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. This work focuses on the study of various data representation techniques its advantages and limitations.**

***Keywords*: Categorical data Representation, Clustering, Unsupervised Learning, K-means.**

1. 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 have 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. At the end conclusion is stated.

2. 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. [2].

In the One Hot Encoding [3] 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 [4], 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) [5], latent Dirichlet allocation (LDA) [6] 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 [7] technique is used to find object to object coupling information based on the conditional probability. The DIstance Learning for Categorical Attributes (DILCA) [8] 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) [9] 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.

Coupled attribute similarity COS [10] technique tries preserves interaction within an attribute, inter-coupled interaction among attributes and the integration of both. It uses occurrence frequency and co-occurrence frequency for calculating intra-coupled interaction and inter coupled interaction respectively.

To overcome the limitation of existing approaches a CURE framework [1] 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.

3. Coupled Data Embedding (CDE)

The system uses coupled unsupervised categorical data representation (CURE) framework for numerical data representation of categorical dataset. The coupled data embedding CDE algorithm follows the CURE framework and generates numerical representation of categorical data.

In CDE, it initially captures the coupling among majority values based on the value coupling. For value coupling it initially generates occurrence-based Value Influence Matrix and Co-occurrence-based Value Influence Matrix. Then the system learns value clusters with different granularity of k means algorithm. The value of k is gradually increased based on the proportion factor.

Then it learns value to value cluster affiliation. The CDE algorithm removes the clusters with less discriminative information. Using the filtered value clusters CDE learns the couplings between the value clusters with PCA. The less important attributes are removed using PCA projection technique. Based on the generated matrix object representation is generated by concatenation of value representation.

A. CDE Algorithm

Input: D - data set,

A. Proportion factor
B. Dimension reducing factor

Output: the numerical representation of objects

Processing:

1. Generate Mo and Mc Matrix
2. Initialize CI = EMPTY
3. For matrices Mo and Mc
4. Initialize cluster count = 2
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. Calculate Z matrix as: Z = CI mean(CI)
12. S: Find covariance matrix of Z
13. Generate Y matrix from SVD(S)
14. Calculate V as ZY T
15. Remove the columns from V whose range is less than B
16. Generate O by the concatenation of V

B. CDE vs Other Encoding Methods [1]:

1. Accuracy: The CDE algorithm is compared with other data encoding methods like 0-1, 0-1P and IDF. After data encoding, clustering is performed using k-means clustering algorithm and the performance of various encoding method is evaluated using Fscore. The test is conducted on various UCI repository categorical datasets like Soybeansmall, Dermatology, adult, Lymphography, Zoo, Mushroom, etc, [11]. The CDE algorithm has significant improvement in FScore as 9%, 5% and 19% with respect to 0-1, 0-1P and IDF encoding methods. The CDE generates slightly higher Fscore than COS, DILCA and ALGO encoding methods.
2. Time: The CDE execution is slower than 0-1, 0-1P and IDF encoding but faster than COS, DILCA and ALGO.

4. Limitations of CDE

The CURE framework mainly focuses on value cluster learning. For value cluster learning CDE algorithm follows the strategy of K means clustering. Hence the limitation of k means algorithm like detecting the special shape of clusters and overlapping clusters also applicable for CDE algorithm.

5. Conclusion

Categorical data cannot be directly processed by machine learning algorithm. The categorical data need to be converted in to numerical format. For good numerical 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. CDE algorithm follows the CURE framework. It uses PCA technique to reduce dimension count but it does not preserve the pair wise distance between data points. A new technique should be developed that preserve intrinsic property of data in dimensionality reduction.

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