CLUSTERING APPROACH USING HIERARCHICAL COUPLING LEARNING FOR CATEGORICAL DATA

1Nilam Patil & 2Prof. S.M.Kamalapur

1ME Student, 2Professor
Department of Computer Engineering,
K.K. Wagh Institute of Engineering Education and Research, Nashik
Savitribai Phule Pune University, Maharashtra, India

Abstract: 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 this work, Updated Coupled Data Embedding (UP-CDE) algorithm is proposed to generate of numerical representation of categorical data. This algorithm is based on Coupled Unsupervised categorical data Representation (CURE) framework. It uses hierarchical learning structure. It defines value to value as well as value clusters coupling. In UP-CDE unwanted features are removed at early stage as compared to the existing CDE algorithm without hampering the accuracy. For early stage feature removal the Principal component analysis (PCA) is used. The clusters of numerical data are generated using k-means algorithm and accuracy is evaluated.

IndexTerms - Categorical data Representation, clustering, Unsupervised Learning, k-means.

I. 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.

II. 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. PCA is used to transfer the data in PC space. In PC space, the data matrix is represented as a new matrix with linearly uncorrelated components.

**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.

**Limitations Of CDE:**

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

**III. 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. This technique converts a data matrix with linearly correlated variables to a new matrix with linearly uncorrelated components. PCA also supports for dimension reduction. There is need to develop a system, to reduce dimension and generate new dimension set by preserving only important attributes.

**IV. SYSTEM DETAILS**

4.1. Preliminaries

4.1.1. Joint Probability

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

\[ P(v_i, v_j) = \frac{\# \{ V_i = v_i \land V_j = v_j \mid x \in X \} }{N} \]  (1)

F is a feature in dataset and \( V_i \) and \( V_j \) are two values of instance \( x \) of features \( i \) and \( j \).

\( N \) : total number of instances.

4.1.2. Normalized mutual information:

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

\[ \Psi(f_a, f_b) = -\sum_{v_i \in V_a} \sum_{v_j \in V_b} p(v_i, v_j) \log \frac{p(v_i, v_j)}{h(f_a) + h(f_b)} \]  (2)

Where, \( p(v_i) \) is the occurrence frequency of \( v_i \)

\[ h(f_a) = -\sum_{v_i \in V_a} p(v_i) \log p(v_i) \]  (3)

© 2020 IJRAR August 2020, Volume 7, Issue 3 www.ijar.org (E-ISSN 2348-1269, P-ISSN 2349-5138)
4.1.3. Occurrence-based Value Influence Matrix

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

$$M_o = \begin{bmatrix} \phi_o(v_{i1}, v_{j1}) & \cdots & \phi_o(v_{i1}, v_{jL}) \\ \vdots & \ddots & \vdots \\ \phi_o(v_{iL}, v_{j1}) & \cdots & \phi_o(v_{iL}, v_{jL}) \end{bmatrix}$$  \hspace{1cm} (4)

Where,

$$\phi_o(v_i, v_j) = \psi(f_i, f_j) \times \frac{P(v_i)}{P(v_j)}$$

is the coupling function.

4.1.4. Co-occurrence-based Value Influence Matrix

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

$$M_c = \begin{bmatrix} \phi_c(v_{i1}, v_{j1}) & \cdots & \phi_c(v_{i1}, v_{jL}) \\ \vdots & \ddots & \vdots \\ \phi_c(v_{iL}, v_{j1}) & \cdots & \phi_c(v_{iL}, v_{jL}) \end{bmatrix}$$  \hspace{1cm} (5)

Where,

$$\phi_c(v_i, v_j) = \frac{P(v_i, v_j)}{P(v_i)}$$

is the coupling function

4.2. 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 updated coupled data embedding (UP-CDE) algorithm is proposed. This algorithm follows the CURE framework and generates numerical representation of categorical data.

As per the CURE framework, CDE[1] initially focuses on value coupling feature of dataset. For value coupling, occurrence-based Value Influence Matrix $M_o$ and Co-occurrence-based Value Influence Matrix $M_c$ matrices are generated. After finding the value coupling the value clusters are performed. For value cluster generation, K-means 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 K-means 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. Too many attributes are generated to represent a single value. The attributes with less information are removed using linear transformation method: PCA and matrix $V$ is generated. Using this technique the dimensions of dataset are reduced and attributes which provide more discriminative information are preserved.

The Principal Component Analysis (PCA) is the projection technique. In this technique the complete dataset is projected on different space This technique helps to clean the dataset to make it easy for further processing. The dimension reduction is depending on the user input. User need to input the dimension reduction rate for generation of $V$ matrix. In CDE the attributes are further removed whose range is less than the predefined $\beta$ threshold.

The generated matrix $V$ is contains $r$ rows and $d$ dimensions. Each row represents a unique database value with $d$ dimension. The final data representation matrix $O$ is generated by concatenating the rows in matrix $V$. This is the numerical representation of data. A clustering using K-means algorithm is applied on this dataset and F-score is evaluated. In the following section detailed steps of algorithm CDE and LEE are given.

4.3. Algorithm

Algorithm 1: UP-CDE

Input:

$D$ - data set, $d$ - reduced dimension count $\alpha$ - proportion factor $\beta$ - Reduction Factor

Output:

$O$: the numerical representation of objects

Processing:

1. Generate $M_o$ and $M_c$ Matrix using preliminary 3 and 4.
2. For matrices $M_o$ and $M_c$
3. Initialize cluster count = 2
4. Initialize $CI = EMPTY, CS = EMPTY$
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($CS$)/$k$ $\geq \alpha$
11. $V$: Apply PCA($Cl, d$) Algorithm
12. Remove Columns from $V$ whose range is less than $\beta$
13. $O$: Generate representation Matrix by concatenating $V$
14. Return $O$
Algorithm 2: PCA:

**Input:** CI - data set, d: reduced dimension count

**Output:**
V : Modified dataset with dimension

**Processing:**
1. M: Find mean of each attribute and generate mean vector
2. CI': Normalize data by subtracting mean from each value
3. Cov: Find Covariance matrix
4. Find Eigen values($\lambda_1, \lambda_2, .., \lambda_n$) and eigen vectors ($e_1,e_2,...,en$) for n dimensions
5. Sort eigen values in descending order
6. Find top d eigen vectors base on top d eigen values
7. Generate matrix W with dimension n*d
8. V: $(CI \times W^T)$ Transform sample in CI in new Space using multiplication
9. Return V

V. RESULT AND DISCUSSION

5.1. Experimental Setup
The hardware and software requirements are mentioned in table 1 for system implementation and testing.

| **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 |

5.2. Datasets
UCI[11] benchmark data sets datasets are used for system testing. Following table gives the detailed description of dataset.

<p>| <strong>Table 2 : Dataset Description</strong> |
|-----------------|------------------|</p>
<table>
<thead>
<tr>
<th><strong>Sr. No.</strong></th>
<th><strong>Dataset</strong></th>
<th><strong>Number of Attribute</strong></th>
<th><strong>Number of instances</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mashroom</td>
<td>22</td>
<td>8124</td>
</tr>
<tr>
<td>2</td>
<td>Tumor</td>
<td>17</td>
<td>339</td>
</tr>
<tr>
<td>3</td>
<td>Soybean-small</td>
<td>35</td>
<td>47</td>
</tr>
<tr>
<td>4</td>
<td>Lymphography</td>
<td>18</td>
<td>148</td>
</tr>
</tbody>
</table>

5.3. Performance Measures

5.3.1. F-score
F-score is evaluated after applying the K-means clustering algorithm on numerical representation dataset. The clustering result is compared with the class values. Let A be the class distribution result and B is the clustering result then the F-score can be calculated using precision and recall values. For this initially cluster to class mapping is performed. For evaluating precision and recall following four values are extracted:
true positive: number of elements matching with class and its cluster number
false negative: number of elements matching with class but not the cluster number
false positive: number of elements from different class belonging to same cluster
true negative: number of non matching class elements that do not belonging to same cluster
Using these values precision can be calculated as:

$$Precision = \frac{true \ positives}{true \ positives + false \ positives}$$  (6)
Recall can be calculated as:

\[ \text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \] (7)

And F-score is calculated as:

\[ \text{F-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \] (8)

### 5.3.2. Evaluation Time

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

### 5.4. Results

The system applies PCA with different reduction rate. The attributes are reduced using PCA technique. If reduction rate is defined as 1 then no attribute is removed. While user selects the reduction rate as 0.8, 0.6, 0.4, 0.2 then only 80%, 60%, 40% and 20% of attributes are preserved in dataset respectively. These attributes are the projection of all other attributes. In CDE, one more attribute filter is added with respect to the range of values it covers. The range of values is less than the predefined threshold value \( \beta \) then those attributes are removed. The finally generated attribute count in matrix \( V \) is same even we vary the reduction rate. The attributes those are not get filtered in PCA are removed in range filter. The value of \( \beta \) is set to 0.0000000001.

#### Table 3: Dimension Analysis

| Dataset       | Dimension Count \( |V| \) (R-rate:1.0) | Dimension Count \( |V| \) (R-rate:0.8) | Accuracy (R-rate:0.6) | Dimension Count \( |V| \) (R-rate:0.4) | Dimension Count \( |V| \) (R-rate:0.2) |
|---------------|-------------------|------------------|-----------------|------------------|------------------|
| Soybean-small | 55                | 55               | 55              | 55               | 52               |
| Mashroom      | 93                | 93               | 93              | 93               | 93               |
| Tumor         | 32                | 32               | 32              | 32               | 32               |
| Lymphography  | 56                | 56               | 56              | 56               | 56               |

#### Time Evaluation

The time required for processing is captured in milliseconds. The system is tested for various dataset. The time is captured with various reduction rate values. As we decrease the reduction rate value, the time required for processing also decreases. With minimum reduction count the system efficiency is improved.

#### Table 4: Time Analysis

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Time in MilliSec (R-rate:1.0)</th>
<th>Time in MilliSec (R-rate:0.8)</th>
<th>Time in MilliSec (R-rate:0.6)</th>
<th>Time in MilliSec (R-rate:0.4)</th>
<th>Time in MilliSec (R-rate:0.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soybean-small</td>
<td>4279</td>
<td>4064</td>
<td>3960</td>
<td>3718</td>
<td>3137</td>
</tr>
<tr>
<td>Mashroom</td>
<td>25007</td>
<td>23816</td>
<td>23504</td>
<td>22886</td>
<td>22433</td>
</tr>
<tr>
<td>Tumor</td>
<td>2164</td>
<td>2109</td>
<td>2065</td>
<td>1985</td>
<td>1833</td>
</tr>
<tr>
<td>Lymphography</td>
<td>4133</td>
<td>3872</td>
<td>3634</td>
<td>3336</td>
<td>3011</td>
</tr>
</tbody>
</table>

The following figure show the time analysis in bar chart format. The time required for processing is captured for each dataset. The time required for processing is depend on the dataset size. Mashroom dataset contains too many instances and hence required higher time as compared to the other datasets. The time required for processing with different reduction rate is compared in this chart. As we decrease the reduction rate count the time required for processing also decreases. The number of attributes in dataset is decreases in PCA at early stage and it helps to reduce the further processing. This improves the efficiency of system.
Accuracy Evaluation

The accuracy of dataset is evaluated using k-means clustering algorithm. After creating the numerical representation of data, K-means algorithm is applied on dataset. Using K-means algorithm the clusters are created. The clustering count is defined as per the class count in dataset. The clustering result is compared with the class values and accuracy of clustering is evaluated. Following table 5 shows the accuracy comparison with different reduction rate values. The accuracy of the system remains same even we change the reduction rate.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy (R-rate:1.0)</th>
<th>Accuracy (R-rate:0.8)</th>
<th>Accuracy (R-rate:0.6)</th>
<th>Accuracy (R-rate:0.4)</th>
<th>Accuracy (R-rate:0.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soybean-small</td>
<td>0.88189</td>
<td>0.88189</td>
<td>0.88189</td>
<td>0.88189</td>
<td>0.88189</td>
</tr>
<tr>
<td>Mashroom</td>
<td>0.666943</td>
<td>0.666943</td>
<td>0.666943</td>
<td>0.666943</td>
<td>0.666943</td>
</tr>
<tr>
<td>Lymphography</td>
<td>0.689382</td>
<td>0.689382</td>
<td>0.689382</td>
<td>0.689382</td>
<td>0.689382</td>
</tr>
</tbody>
</table>

The flowing graph shows the accuracy comparison among clustering results derived from datasets generated using different reduction rate. The accuracy does not vary with respect to the reduction rate used in CDE algorithm. The results are taken for 3 different datasets with 5 different reduction rate count.

5.5. Observations

The PCA technique reduces the dimension at the early stage as compared to the CDE[1] algorithm and hence further processing overhead of attribute values is get reduced. This leads to improve the efficiency of system.

- During the evaluation, we notice that our proposed approach,
- Generate same dimensions as Existing CDE technique.
- Accuracy of system equal to the existing system
- Processing time is get reduced.
- Less time required for minimum reduction rate.
- Lower reduction rate do not hamper the accuracy.

VI. CONCLUSION AND FUTURE WORK

In our research attempt, we have focused on the numerical representation of categorical data with limited attributes using CURE framework. For good data representation intrinsic data characteristics should be effectively captured. The CURE framework uses hierarchical value couplings using value to value interaction and cluster –correlation. The up-CDE algorithm follows the CURE framework and generates numerical representation of data. In UP-CDE unwanted features are removed at early stage as compared to the existing CDE algorithm. PCA technique is used to removing features. This improves the system efficiency without hampering the accuracy. The average time reduced for all processing is 20% for 0.2 reduction rate as compared to the existing CDE technique. For accuracy evaluation K-means clustering is used.

In future system can be implemented using complex coupling among attribute values i.e. more than two values can be considered for value coupling information extraction.
REFERENCES
