Supplementary Material of Machine Learning Integrated Credibilistic Semi Supervised Clustering for Categorical Data

Jnanendra Prasad Sarkar\textsuperscript{a,c,d}, Indrajit Saha\textsuperscript{b,d,}\textsuperscript{∗}, Sinjan Chakraborty\textsuperscript{c,d}, Ujjwal Maulik\textsuperscript{c}

\textsuperscript{a}Larsen & Toubro Infotech Ltd., Pune, India
\textsuperscript{b}Department of Computer Science and Engineering, National Institute of Technical Teachers’ Training and Research, Kolkata, India
\textsuperscript{c}Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
\textsuperscript{d}Equally contributed

Abstract

In real life, availability of correctly labeled data and handling of categorical data are often acknowledged as two major challenges in pattern analysis. Thus, clustering techniques are employed on unlabeled data to group them according to homogeneity. However, clustering techniques fail to make a decision while data are uncertain, ambiguous, vague, coincidental and overlapping in nature. Hence, in this case, the use of semi supervised technique can be useful. On the other hand, real life datasets are majorly categorical in nature, where natural ordering in attribute values is missing. This special property of categorical values with the inherent characteristics like uncertainty, ambiguity and vagueness makes clustering more complicated than numerical data. In recent times, credibilistic measure shows better performance over fuzzy and possibilistic measures while considering similar inherent characteristics in numerical data. Thus, these facts motivated us to propose a semi supervised clustering technique using credibilistic measure with the integration of machine learning techniques to address the above mentioned challenges of clustering categorical data. This semi supervised technique first clusters the dataset into $K$ subsets with the proposed Credibilistic $K$-Mode, where credibilistic measure helps to determine the homogeneity by avoiding coincident clustering problem as well as finds the points those are certain to the clusters. Thereafter, in the second part of the semi supervised technique, clustered dataset is used to build a supervised model for classification of other unlabeled or uncertain data. This technique not only handles the unlabeled data better, but also yields improved results for uncertain or ambiguous data e.g, if the credibilistic measure is same for a data point in multiple classes. The results of the proposed technique are demonstrated quantitatively and visually in comparison with widely used state-of-the-art methods for eight synthetic and four real life datasets. Finally, statistical tests have been conducted to judge the statistical significance of the results produced by the proposed technique.

Keywords:
Categorical data, Credibilistic clustering, Friedman test, Fuzzy set, Machine learning, Possibilistic measure, Semi supervised clustering, Statistical significance.
1. Brief description of machine learning methods

Among different machine learning (ML) methods, $K$ Nearest Neighbor (K-NN) [1] classifies data based on the $K$ nearest neighbors from the training dataset. This method does not make any assumption on the underlying data distribution. The number, $K$, decides how many neighbors influence the classification. Usually it is kept as odd in case the number of classes is two.

Support Vector Machine (SVM) is another widely used [2] supervised method, which is originally developed by Vapnik [3]. SVM constructs a separating hyperplane [4] in $d$-dimensional space to separate the input dataset into two classes. The hyperplane maximizes the margin between two classes. SVM method is primarily designed for two class problems. This can be extended for multiclass problems by designing few one-against-all or one-against-one [5] two class SVMs. SVM transforms the input dimension into high dimensional feature space for linearly non-separable input dataset. Thereafter, it constructs a linear hyperplane for classification. The process is computationally complex. SVM tackles this type of problem by defining a suitable \textit{Kernel} function, where the problem is addressed only in the input space.

While dealing with high dimensional data, K-NN is not efficient without approximate dimension reduction or feature selection. SVM can be useful to deal with high dimensional data. However, it is not robust when there are large number of irrelevant variables [6]. Hence, feature selection also became an important factor. Moreover, it is difficult to ensure the better accuracy in case of multiclass problem.

Artificial Neural Network (ANN) [7] is also a popular, efficient and widely used machine learning method which is a mathematical model to mimic human nervous system. The basic processing element of the model is called artificial neuron or simply neuron. The synapses of a neuron are designated by weight of the connection which modulates the effects of input signals. ANN can be used for datasets which are not linearly separable. The nonlinear characteristic is represented in ANN by transfer function. The weighted sum of input signals are considered as neuron impulse and it is transformed by transfer function. The learning capability is tuned by adjusting the weights. ANN is typically used for complex classification problem.

On the other hand, Decision Tree (DT) [8] is such a tree based method which builds classification model as a form of tree structure. The tree is incrementally developed by partitioning the dataset into subsets based on information gain at each node. Intermediate nodes are called decision node whereas, leaf nodes represent classification. Random Forest (RF) [9] is also developed to handle the same issues in form of ensemble. This method combines a forest of decision trees on random input vectors and with the same distribution for all trees in the forest. Thereafter, it splits the nodes on random subset of features. RF can be applied for three purposes, such as classification, clustering and regression.

*Corresponding author: indrajit@nitttrkol.ac.in*
2. Cluster Validity Indices

The results of various clustering techniques for the same input dataset can vary, because input parameters of any technique substantially affect the overall behavior and execution of the technique. The main objective of a cluster validity index [10–16] is to validate clustering solution. Following subsections briefly describe the cluster validity measures which have been used in article.

2.1. Minkowski Score

Minkowski Score (MS) [10] is a measure of the quality of a solution as compared to true clustering. Let $T$ be the “true” solution and $S$ the solution we want to measure. Moreover, $n_{11}$ is the number of pairs of elements that are in the same cluster of both $S$ and $T$. Similarly, $n_{01}$ and $n_{10}$ represent the number of pairs that are in the same cluster only in $S$, and the number of pairs that are in the same cluster in $T$ respectively. Minkowski Score is then defined as:

$$MS = \sqrt{\frac{n_{01} + n_{10}}{n_{11} + n_{10}}}$$  \hspace{1cm} (1)

Minimum value of MS is 0, while lower value denotes better clustering.

2.2. Percentage of Correct Pair

Percentage of Correct Pair ($\%CP$) [11] is defined as below

$$\%CP = \frac{\text{number of pairs correctly clustered into the same cluster}}{\text{pairs actually in the same cluster}}$$  \hspace{1cm} (2)

Higher value of CP signifies better clustering. It gives the result in percentage form. Therefore, 100% means perfect clustering.

2.3. Adjusted Rand Index

Adjusted Rand Index (ARI) [12] represents a relation between true cluster and the computed cluster of the dataset. If $T$ and $C$ are the true clustering of the dataset and the computed clustering through some clustering technique, then ARI can be defined as

$$ARI(T, C) = \frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)}$$  \hspace{1cm} (3)

where, $a$, $b$, $c$, and $d$, respectively, denote the number of pairs belonging to the same cluster of both $T$ and $C$, the number of pairs belonging to the same cluster of $T$ but to different clusters of $C$, the number of pairs belonging to different clusters of $T$ but to the same cluster of $C$ and the number of pairs belonging to different clusters of both $T$ and $C$. $ARI$ value varies between zero and one. Higher value indicates that the resulted clustering is more close to the actual one.
2.4. Xie-Beni Index

Xie-Beni (XB) [16] index is the ratio of the total fuzzy cluster variance, $\sigma$, to the minimum separation, $\zeta$, of the clusters. If total number of data points is $n$, then XB can be defined as below.

$$XB = \frac{\sigma}{n \times \zeta}$$  \hspace{1cm} (4)

where

$$\sigma = \sum_{l=1}^{K} \sum_{i=1}^{n} (\mu_{il})^{2} D(v_{l}, x_{i})$$  \hspace{1cm} (5)

and

$$\zeta = \min_{h \neq l} \{D(v_{h}, v_{l})\}$$  \hspace{1cm} (6)

$D(v_{l}, x_{i})$ is the dissimilarity measure between cluster mode $v_{l}$ and data point, $x_{i}$.

3. Dissimilarity measure

The dissimilarity measure which is used in article is described in this section. Let two categorical data points described by $m$ categorical attributes are $x_{i} = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$, and $x_{j} = \{x_{j1}, x_{j2}, \ldots, x_{jm}\}$. The distance measure between $x_{i}$ and $x_{j}$, $D(x_{i}, x_{j})$ can be defined as follows:

$$D(x_{i}, x_{j}) = \sum_{k=1}^{m} \delta(x_{ik}, x_{jk})$$  \hspace{1cm} (7)

where

$$\delta(x_{ik}, x_{jk}) = \begin{cases} 
0, & \text{if } x_{ik} \neq x_{jk} \\
1, & \text{if } x_{ik} = x_{jk}
\end{cases}$$  \hspace{1cm} (8)

4. Partial experimental results

The comparative study of integrated methods on average %CP [11] and ARI [12] values are reported in Figs. 1 and 2. It has been observed that CrKMd-RF produces better results consistently for different synthetic and real life datasets used for our experiments.

5. Partial visual results

To visualize a clustering solution using visual assessment of tendency (VAT) [17], first the data are reordered according to the class labels given by the solution. Thereafter the distance matrix is computed on this reordered data matrix. In the graphical plot of the distance matrix, the boxes lying on the main diagonal represent the clustering structure. VAT plots of datasets Cat_250,15,5, Cat_300,8,3, Cat_300,15,5, Cat_500,20,10, Dermatology and Mushroom are shown in Figs. 3 - 8.
Figure 1: Average values of %CP for CrKMd integrated ML methods for different datasets

Figure 2: Average values of ARI for CrKMd integrated ML methods for different datasets

Figure 3: VAT plot of (a) True clusters and the clusters produced by (b) FKMd (c) CrKMd (d) MLCrKMd for Cat_250_15_5 dataset

Acknowledgment

This work was partially supported by the grants from the Department of Science and Technology, India (DST/INT/POL/P-36/2016).
Figure 4: VAT plot of (a) True clusters and the clusters produced by (b) FKMd (c) CrKMd (d) MLCrKMd for Cat300_8_3 dataset

Figure 5: VAT plot of (a) True clusters and the clusters produced by (b) FKMd (c) CrKMd (d) MLCrKMd for Cat300_15_5 dataset

Figure 6: VAT plot of (a) True clusters and the clusters produced by (b) FKMd (c) CrKMd (d) MLCrKMd for Cat500_20_10 dataset

References

Figure 7: VAT plot of (a) True clusters and the clusters produced by (b) FKMd (c) CrKMd (d) MLCrKMd for Dermatology dataset

Figure 8: VAT plot of (a) True clusters and the clusters produced by (b) FKMd (c) CrKMd (d) MLCrKMd for Mushroom dataset
