Categorical Clustering By Converting Associated Information

Dongmin Cai  Stephen S-T Yau

Abstract—Lacking an inherent “natural” dissimilarity measure between objects in categorical dataset presents special difficulties in clustering analysis. However, each categorical attributes from a given dataset provides natural probability and information in the sense of Shannon. In this paper, we proposed a novel method which heuristically converts categorical attributes to numerical values by exploiting such associated information. We conduct an experimental study with real-life categorical dataset. The experiment demonstrates the effectiveness of our approach.

Keywords—Categorical, Clustering, Converting, Information

I. INTRODUCTION

GROUPING a set of objects into classes of similar objects is one of the major topics [3], [16], [18], [19]. A large number of clustering algorithms exist in the literature. In general, they can be classified into several categories [16], i.e. partitioning methods [20], [22], hierarchical methods [3], [7], [18], density-based methods [2], [8], [15], model-based methods [23] [9], and etc.

It is natural to approach this problem by computing the similarities between objects when distance function is naturally defined [16]. Numerical data has such geometric properties. Unfortunately, problems arise when it comes to categorical clustering, i.e. clustering data whose domains are discrete and not naturally ordered. It is hard to define which attribute is “smaller” or “bigger”. Lacking an inherent “natural” order makes a large number of traditional similarity measures ineffective. In the field of data mining, a lot of categorical clustering algorithms exist in the literature. In [12], the authors present a novel method which exploits information associated with each categorical attribute in a given dataset. Using such information, we are able to heuristically convert categorical attributes to numerical values. Our approach is derived from the concept of information theory and probabilistic model. It neither simply encodes categorical attributes, which does not necessarily produce reasonable results, nor sets entropy criterion for partitioning.

In a thorough real dataset evaluation, we demonstrate the effectiveness of our approach.

The rest of the paper is organized as follows. Section II discusses related work on clustering categorical data. Section III describes cluster problem from mathematical aspect. Section IV introduces some definitions used in this paper. Section V expands discussion about the definitions. Section VI heuristically establishes converting model for categorical data. Section VII presents real-life experiment results. Section VIII follows summary.

II. RELATED WORK

In the standard clustering methods, COBWEB [9] is one of the popular algorithms for categorical data. It utilizes incremental learning and dynamically builds a dendrogram. Its conceptual learning property can describe cluster intrinsically. The technique can be extended to transactional systems. The application can be extended to transactional systems. The algorithm employs links not distances when merging clusters. From functional analysis, [12] presents STIRR. The techniques can be studied analytically in terms of certain types of non-linear dynamic systems. The application can be extended to transactional data. For modification, [30] proposes a revised dynamic system to solve convergence problem in [12].

In the area of categorical clustering, researchers are still looking for the potential relationship between categorical attributes and numerical values and making effort to employ numerical criteria for clustering. CACTUS [11] is such an algorithm. It defines inter-attribute and intra-attribute summaries which tell us how well values from different and the same attribute are related respectively. In [26], the author mentions an acceptable approach to use K-means algorithm to cluster categorical data. It converts multiple categorical attributes into binary attributes without order.
Information-related methods are attracting researcher attentions recently. They observe that clusters of similar objects have lower entropy. In [4], the authors design a heuristic algorithm, COOLCAT, which is capable of efficiently clustering large data sets with categorical attributes. [6] presents a solution which provides the dissimilarities between objects based on an information-theoretical definition of dissimilarities between partitions of finite sets. Using entropy as criterion, [21] proposes to search the partitions to minimize the information-based criterion. In fact, each categorical attributes from a given dataset provides natural dissimilarities between partitions of finite sets. Using entropy prov ides the information and uncertainty of a set of possible events. Therefore, entropy is independent on attribute index $i$. From now on, we denote $n_k^i = n_k^1 = n_k^2 = \ldots = n_k^m$, then the number of objects in the dataset $D$ is

$$n = \sum_{i=1}^{m} n_k^i$$

Now we are ready to introduce some definitions which are used to measure the information for categorical attribute. In Shannon’s pioneer work [27], the entropy is defined to measure how much “choice” is involved in the selection of the event or of how uncertain we are of the outcome. In other words, entropy provides the information and uncertainty of a set of possible events. Therefore, $E(r^i_j)$ is used to measure the amount of information needed for the set of objects having $c_i = r^i_j$ to be converted to target attribute.

As the opposite, we introduce Weight of Certainty and Information (WCI) to measure the weight of certainty and information for object having $c_i = r^i_j$. It has the property that the greater WCI is, the more information and certainty the object having $c_i = r^i_j$ brings.
Definition 4.4: Let $WCI(r_j^i)$ be the Weight of Certainty and Information (WCI) for objects having $c_i=r_j^i$ to be converted to target attribute for $1 \leq i \leq m$ and $1 \leq j \leq a_i$.

$$WCI(r_j^i) = \frac{\log_2 l - E(r_j^i)}{\log_2 l}$$ (9)

Shannon has shown $0 \leq E(r_j^i) \leq \log_2 l$ in his work [27]. It immediately follows that $0 \leq WCI(r_j^i) \leq 1$, where (a) $WCI(r_j^i)=0$ if and only if $p_{j1}^i=p_{j2}^i=...=p_{jr}^i=1/l$; (b) $WCI(r_j^i)=1$ if and only if $p_{j1}^i=1$ for some $k=s$ ($1 \leq s \leq l$) and all other $p_{jk}^i$, for $k \neq s$ are 0.

Definition 4.5: Let $IGR(A_i)$ be the Information Gain Ratio (IGR) for categorical attribute $A_i$ for $1 \leq i \leq m$.

$$IGR(A_i) = -\sum_{i=1}^{a_i} n_i \log \frac{n_i}{n} - \sum_{j=1}^{a_i} q_j E(r_j^i)$$

$$\quad - \sum_{j=1}^{a_i} q_j \log_2 q_j$$ (10)

here

$$q_j = \frac{n_j}{n}$$ for $1 \leq i \leq m$ and $1 \leq j \leq a_i$ (11)

Quinlan names information gain as the measure of information for an attribute [24]. For example, in the construction of decision tree, the higher information gain an attribute $A_i$ is, the greater information $A_i$ provides. However, Quinlan realizes that information gain has a strong bias for attributes with many values. To remove such bias, Quinlan addresses that IGR is a normalization that reduces the bias [25].

V. DISCUSSION

In this section, let us discuss the terms defined in section IV. Like [11], the method defines inter-attribute and intra-attribute summaries which tell us how well values from different and the same attribute are related respectively, our approach evaluates information provided by attributes from those two aspects.

(1) $p_{jk}^i$

From (8), we know that $p_{jk}^i$ provides natural probability that an object having $c_i=r_j^i$ is converted to target attribute with value $t_k$. For example, $p_{jk}^i=0.31$ heuristically tells us that if $c_i=r_j^i$ there is $31\%$ chance that the object is mapped to $t_k$.

(2) $WCI(r_j^i)$

$p_{jk}^i$ provides converting information of object having $c_i=r_j^i$ from probability concept. Besides the probability, there is WCI associated with $r_j^i$. The bigger WCI($r_j^i$) results in, the more certainty (weight) we are sure about the converting. For example, if $r_j^i$ can fully tell which target value an object is converted to, i.e. $t_k$ (where $p_{jk}^i=1$ and $p_{jk'}^i=0$ for all $k \neq r$), then $r_j^i$ produces the most certainty. By Def 4.4, the weight $WCI(r_j^i)=1$; if $r_j^i$ has no preference to any target attribute value, the associated $p_{jk}^i$’s for all $k=1...l$ have equal-probability value $1/l$. Therefore $r_j^i$ produces the most uncertainty and cannot add any information for decision. By Def. 4.4, the weight $WCI(r_j^i)=0$ is used to represent such uselessness.

(3) $IGR(A_i)$:

As we stated before, both $p_{jk}^i$ and $WCI(r_j^i)$ are related to intra-attributes, i.e. attribute $A_i$. However we also need to measure the relevance of different attributes. Thus, an inter-attribute should be measured. In Def. 4.5, we mentioned that $IGR(A_i)$ is used to measure the amount of information that $A_i$ can provide.

We have analyzed the roles that $p_{jk}^i$, $WCI(r_j^i)$, and $IGR(A_i)$ play from information concept. $p_{jk}^i$ can heuristically provide probability that an object having $c_i=r_j^i$ is converted to target value $t_k$. But it’s not completed since we need to know the confidence for such probabilistic converting. The term $WCI(r_j^i)$ fulfills such task to measure the certainty weight. However we realize these two terms are only related to the same attribute. As a measure of information gain for one attribute, $IGR(A_i)$ is added as a weight to measure the difference between attributes.

VI. CONVERTING CATEGORICAL DATA

From the above discussion, we are ready to deploy our proposed method to convert categorical attributes and categorical data.

Definition 6.1: Let $r_j^i$ for $j=1,...,a_i$ be an element in the attribute $A_i$. According to the statement above, $r_j^i$ is converted to a 1-by-$l$ vector

$$v(r_j^i) = [IGR(A_i)WCI(r_j^i)p_{jk}^i]_{k=1..l}$$ (12)
Definition 6.2: Let $O = (c_1, c_2, \ldots, c_m)$ be an object in the set $D$, where $c_i = r_j^i$ for $i = 1, \ldots, m$, and $j = 1, \ldots, a_i$. Assume attributes are independent. Then from Def. 6.1, $O$ is converted to a vector,

$$V = \sum_{i=1}^{m} v(c_i = r_j^i)$$  (13)

Definition 6.3 Let $O$ and $O'$ be distinct objects from the set $D$. Where $O = (c_1, c_2, \ldots, c_m)$ and $O' = (c_1', c_2', \ldots, c_m')$. Following Def. 6.2, $O$ is converted to $V$, and $O'$ is converted to $V'$. The pseudo distance between $O$ and $O'$ is defined by using Euclidean distance:

$$d(O, O') = ||V - V'||_2$$  (14)

So far, we have formally constructed the framework of dissimilarity measure between categorical data. In summary, the proposed clustering process involves three phases:

1. Firstly, we need to estimate the information from inter-attributes and intra-attributes. This estimation can be obtained from a domain expert or by using a training set of objects.

2. Secondly, we heuristically convert each attribute using the associated information. Therefore each object in the dataset is converted numerically with reasonable concept.

3. Finally, traditionally clustering algorithms can be exploited effectively.

VII. EXPERIMENTS

A. Performance Measure

In section VI, the dissimilarity between objects is measured by Euclidean distance. Therefore, some traditional clustering methods, i.e. partitioning methods [20], [22] and hierarchical methods [3], [7], [18], can be applied for clustering based on the distance in (14).

In this paper, we use two measures to evaluate the clustering performance, accuracy rate and entropy. Both of them are used as the measure of external quality. External quality let us evaluate how well the clustering is working by comparing the groups produced by the clustering techniques to known classes [28]. If one clustering algorithm performs better than other clustering algorithms on these measures, then we can have some confidence that it is truly the better clustering algorithm for the situation being evaluated [28].

(1) Accuracy rate measure

Suppose that user specifies the number of clusters, i.e. $K$ clusters. In [1], clustering accuracy rate $r$ is defined as

$$r = \frac{\sum_{j=1}^{K} n_j}{n}$$  (15)

where $n_j = \max\{ |n_j^i| \text{ for all } j \}$ where $|n_j^i|$ is the number of objects from cluster $i$ belongs to target class $j$. In other words, $n_j$ is the dominating number of objects in cluster $i$. A high accuracy rate implies that the clusters obtained are “pure”.

(2) Entropy measure

Entropy [27] has been used as a measure of quality of the clusters for a long time. The smaller the entropy is, the better the performance will be.

Suppose there are $l$ classes for target attribute $T$, and user specifies the number of clusters, i.e. $K$ clusters. For each cluster, we compute $p_j^i$ ($i=1\ldots K$ and $j=1\ldots l$), the “probability” that an object from cluster $i$ belongs to target class $j$. Then the entropy of each cluster $i$ for $i=1\ldots K$ is calculated [27],

$$E_i = -\sum_{j=1}^{l} p_j^i \log_2 p_j^i$$  (16)

Let $n_i$ be the number of objects of cluster $i$, and $n$ is the total number of objects of the dataset. Then the weighted entropy for the clustering performance is calculated,

$$E = \sum_{i=1}^{K} \frac{n_i}{n} E_i$$  (17)

B. Real-life Datasets

Because of the limit of space, we present one real-life categorical datasets from UCI Machine Learning Repository [5] to evaluate our approach.

Congressional votes: It is the United States Congressional Voting Records in 1984. The dataset has 435 records, 168 of them are from Republicans, and 267 of them are from Democrats. Each record corresponds to one congress man’s votes on 16 issues (e.g., education spending, immigration). The votes to all of those 16 issues are simplified to “Yes”, “No”, and “?” (neither “Yes”, nor “No”). A classification label of Republican or Democrat is provided with each data record. Thus it can be treated as target attribute. In summary, the dataset contains 16 categorical attributes plus one target attribute.

C. Pre-Processing

In section IV, we assume the dataset $D$ has summaries. The summaries can be obtained from a domain expert who understands the dataset very well and can provide such information as pre-processing. If such expert is not available, we can use a set of training objects which contain target attribute to approximate those summaries. Here we assume this estimation is obtained from:

(1) E-1: Summaries estimated using samples of 60% records from the dataset. In general, 60% sampling will not provide the summaries information requested by our method exactly. Therefore, it exits some error from categorical attributes converted to numerical values.

(2) E-2: Summaries estimated using the entire dataset. Using the entire dataset, we expect the estimation of summaries is relatively accurate compared to 60% sampling.

We compared our approach using these two training datasets with K-means clustering algorithm using coded categorical attributes [26].
D. Experiment Result and Comparison

Congressional votes has a “natural” cluster (Republican / Democrat) number, $K=2$. So, we are more interested in the results when cluster number is set to 2. Table 1 shows such running results.

| TABLE 1 CLUSTERING RESULTS FOR CONGRESSIONAL VOTES AT CLUSTER NUMBER $K=2$ |
|-----------------|-----------------|-----------------|
| Cluster No.     | # of Democrats  | # of Republicans |
| 1               | 42              | 158             |
| 2               | 225             | 10              |

As the table illustrates,
(1) $K$-means algorithms: Cluster 1 has confidence to represent Republican group. However, around 21% of the members are Democrats. Cluster 2 has confidence to represent Democrat group. And 4% of the members are Republicans.
(2) E-1: Cluster 1 has confidence to represent Republican group. Cluster 2 has confidence to represent Democrat group. The expected accuracy rate (0.8897) is higher than that (0.8805) from $K$-means. And the total expected entropy (0.3948) is less than that (0.4781) from $K$-means.
(3) E-2: Cluster 1 has confidence to represent Republican group. Around 10% of the members are Democrats. Cluster 2 has confidence to represent Democrat group. And 3% of the members are Republicans. The expected accuracy rate (0.9425) is higher than that from $K$-means. And the total expected entropy (0.3040) is less than that from $K$-means.

It’s not surprised to observe that E-2 has better performance than E-1 since the summaries estimated from E-2 is more accurate than that from E-1.

For more comparisons, we run the experiment on cluster number $K$ from 3 to 10. Fig. 1 (a) displays comparison results measured by accuracy rate. Fig. 1 (b) displays the results measured by entropy. In Fig. 1(a), we observe that both E-1 and E-2 have higher accuracy rate than $K$-means. Similarly in Fig. 1(b), both E-1 and E-2 obtain less entropy than $K$-means. This experiment verifies that our proposed approach either using estimated summary from 60% samples or from the entire dataset have better performance than $K$-means.

Therefore we have some confidence to claim that it’s better clustering algorithm for the situation being evaluated.

![Fig. 1 (a) Comparison of accuracy for congressional votes](image1)

![Fig. 1 (b) Comparison of entropy for congressional votes](image2)

VIII. CONCLUSION

In this paper, we proposed a novel method which heuristically converts categorical attributes to numerical values by exploiting information which is associated with intra-attributes and inter-attributes. Our approach is derived from the concept of information theory and probabilistic model. It neither simply encodes categorical attributes, which does not necessarily produce meaningful results, nor sets entropy criterion for partitioning. The results from our experimental study with real-life datasets are very encouraging.

REFERENCES


Stephen S-T Yau is a distinguished professor at University of Illinois at Chicago. He became a IEEE Fellow in 2003. He received a Guggenheim Fellowship in 2000.