International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:11, 2007

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 work have been published [3], [11], [12], [14], [17], [23], [26], [30]. Recently, information-related approaches have been studied [4], [6], [21].

In this paper, we proposed 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.

Manuscript received November 15, 2005.

Dongmin Cai is with the Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, IL 60607 USA (phone: 312-996-3049; fax: 312-996-1491; e-mail: dcai5@ uic.edu).

Stephen S-T Yau is with the Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, IL 60607 USA (phone: 312-996-3065; fax: 312-996-3065; e-mail: yau@ uic.edu).

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. [23] presents other conceptual clustering algorithms and produces conceptual descriptions of clusters. People have sought dissimilarity measures by traditional clustering methods, e.g., Gower's similarity coefficient [13] and other dissimilarity measures [10] applied on the hierarchical clustering methods [3], [18]. By using a simple matching dissimilarity measure and replacing means with modes, [17] presents an algorithm, *K*-modes. *K*-modes extends the traditional *K*-means [22] partitioning clustering from numerical data to categorical data. In [14], the authors propose a novel algorithm, ROCK, a concept of links to measure the similarity between objects. 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 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.

III. Mathematical Preliminaries

In this section, we formally define categorical dataset, object, and cluster from mathematical aspect.

A. Categorical Dataset

Definition 3.1: Let $\{A_1, A_2, \ldots, A_m\}$ be a set of categorical attributes which consists of finite set of values. We let $a_i = |A_i|$ be the size of the attribute A_i for i=1,…,m, then A_i is described as a finite set $\{r_j^i \mid j = 1, ..., a_i\}$. Let

$$D = \{(c_1, ..., c_i, \ldots, c_m),\ c_i \in A_i,\ 1 \le i \le m\} \quad (1)$$

be the set of objects, the dataset.

B. An Object from Categorical Dataset

Definition 3.2: Following Def. 3.1, an object O from D is defined as a vector

$$O = (c_1, c_2, \ldots, c_m) \quad (2)$$

C. Categorical Cluster

For clustering problem, objects are grouped into classes of similar objects [16]. The set of such classes is named class target (class attribute) in this paper.

Definition 3.3: Let us assume t is from a class target $T = \{t_k \mid k = 1, ..., l\}$, $l = |T|$. For clustering problem, there is an association between D and T,

$$\{(c_1, ..., c_i, \ldots, c_m, t),\ c_i \in A_i,\ 1 \le i \le m, \text{ and } t \in T\} \quad (3)$$

Each object $\overline{O}$ with target attribute is defined as an element in this association, i.e.

$$\overline{O} = (c_1, c_2, \ldots, c_m, t) \quad (4)$$

IV. Definitions

In this section, we introduces some definitions used in the remaining of the paper.

We have mentioned that our method exploits information associated with each categorical attributes. Such information is estimated based on the following summaries.

Definition 4.1: Let n_{jk}^i for $1 \le j \le a_i$ and $1 \le k \le l$ denotes a summary associated to attribute A_i, where $n_{jk}^i = |D_{jk}^i|$ and $D_{jk}^i = \{(c_1, ..., c_i, \ldots, c_m, t)$: $c_i = r_j^i$ and $t = t_k\}$.

Definition 4.2: Let n_k^i $(1 \le k \le l)$ be a summary that associated to A_i represents the number of objects having $t = t_k$,

$$n_k^i = \sum_{j=1}^{a_i} n_{jk}^i \quad \text{for } 1 \le i \le m \quad (5)$$

Clearly $\bigcup_{j=1}^{a_r} D_{jk}^u = \bigcup_{j=1}^{a_s} D_{jk}^v$ for any $1 \le u, v \le m$, it follows immediately $n_k^1 = n_k^2 = \ldots = n_k^m$. Therefore n_k^i is independent on attribute index i.

From now on, we denote $n_k = n_k^1 = n_k^2 = \ldots = n_k^m$, then the number of objects in the dataset D is

$$n = \sum_{k=1}^{l} n_k \quad (6)$$

Now we are ready to introduce some definitions which are used to measure the information for categorical attribute.

Definition 4.3: Let $E(r_j^i)$ be the Entropy of the set of objects having $c_i = r_j^i$ to be classified (converted) to target attribute for $1 \le i \le m$ and $1 \le j \le a_i$

$$E(r_j^i) = -\sum_{k=1}^{l} p_{jk}^i \log_2 p_{jk}^i \quad (7)$$

here

$$p_{jk}^i = \frac{n_{jk}^i}{\sum_{k=1}^{l} n_{jk}^i} \quad \text{for } 1 \le i \le m,\ 1 \le j \le a_i,\ 1 \le k \le l \quad (8)$$

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, $E(r_j^i)$ is used to measure the amount of information needed for the set of objects having $c_i = r_j^i$ 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 $c_i = r_j^i$. It has the property that the greater WCI is, the more information and certainty the object having $c_i = r_j^i$ 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 $1 \le i \le m$ and $1 \le j \le a_i$

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

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

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

$$IGR(A_i) = \frac{-\sum_{k=1}^{l} \frac{n_k}{n} \log_2 \frac{n_k}{n} - \sum_{j=1}^{a_i} q_j^i E(r_j^i)}{-\sum_{j=1}^{a_i} q_j^i \log_2 q_j^i} \quad (10)$$

here

$$q_j^i = \frac{\sum_{k=1}^{l} n_{jk}^i}{n} \quad \text{for } 1 \le i \le m \text{ and } 1 \le j \le a_i \quad (11)$$

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_r (where p_{jr}^i=1 and p_{jk}^i=0 for all $k \ne r$), 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 $j = 1,\ldots,a_i$ be an element in the attribute A_i. According to the statement above, r_j^i is converted to a 1-by-l vector

$$v(r_j^i) = [\, IGR(A_i) WCI(r_j^i)\, p_{jk}^i \,]_{k=1\ldots l} \quad (12)$$

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 $i = 1,...,m$, and $j = 1,..., a_i$. Assume attributes are independent. Then from Def. 6.1, O is converted to a vector,

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

Definition 6.3 Let O and O' are 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:

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

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

$$\mathrm{r} = \frac{\sum_{i=1}^{K} n_i}{n} \quad (15)$$

where

$n_i = max\{ n_j^i$ for all $j=1\ldots l \mid n_j^i$: number of objects from cluster i belongs to target class $j\}$. In other words, n_i 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…K and j=1...l), the "probability" that an object from cluster i belongs to target class j. Then the entropy of each cluster i for i=1…K is calculated [27],

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

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,

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

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 I
CLUSTERING RESULTS FOR *CONGRESSIONAL VOTES* AT CLUSTER NUMBER K=2

K-means algorithm using coded categorical attributes Accuracy rate =0.8805 ; Entropy =0.4781		
Cluster No.	# of Democrats	# of Republicans
1	42	158
2	225	10

E-1: Converting Method with 60% samples training Accuracy rate = 0.8897; Entropy =0.3948		
Cluster No.	# of Democrats	# of Republicans
1	47	167
2	220	1

E-2: Converting Method with entire dataset training Accuracy rate = 0.9425; Entropy =0.3040		
Cluster No.	# of Democrats	# of Republicans
1	17	160
2	250	8

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.

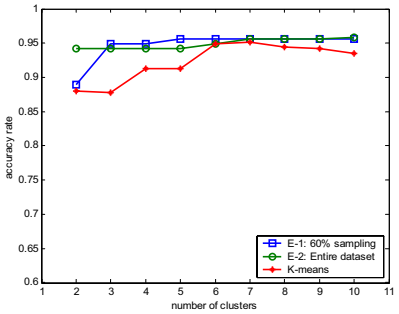

Fig. 1 (a) Comparison of accuracy for *congressional votes*

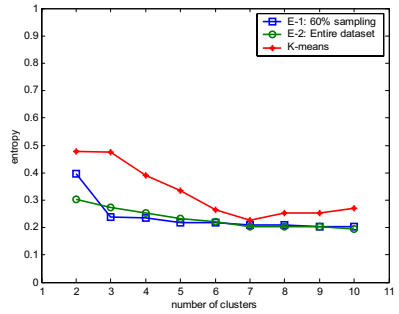

Fig. 1 (b) Comparison of entropy for *congressional votes*

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

[1] C. C. Aggarwal, A human-computer interactive method for projected clustering, IEEE Transactions on Knowledge and Data Engineering, 16(4), 448-460, 2004.
[2] M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. In Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99), pages 49{60, Philadelphia, PA, June 1999.
[3] M.R. Anderberg, Cluster analysis for applications, Academic Press, 1973.
[4] D. Barbara, Y. Li, J. Couto, COOLCAT: An entropy-based algorithm for categorical clustering. In: CIKM Conference. McLean, VA, 2002.
[5] C.L. Blake and C.J. Merz, UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html
[6] D. Cristofor and D. A. Simovici, An information-theoretical approach to clustering categorical databases using genetic algorithms. In Proceedings

of the Workshop on Clustering High-Dimensional Data and Its Applications (SIAM ICDM), pages 37–46, Washington, 2002.
[7] Richard O. Duda and Peter E. Hard, Pattern classification and scene analysi. A wiley-Interscience Publication, New York, 1973.
[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96), pages 226{231, Portland, Oregon, Aug. 1996.
[9] D. Fisher, Improving inference through conceptual clustering. In Proc. 1987 National Conference Artificial Intelligence (AAAI'87), pages 461-465, Seattle, WA, July 1987.
[10] K.C. Gowda and E. Diday, Symbolic clustering using a new dissimilarity measure. Pattern Recognition, 24(6): 567-578, 1991.
[11] V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS: Clustering categorical data using summaries. In ACM SIGKDD Int'l Conference on Knowledge discovery in Databases, 1999.
[12] David Gibson, Jon Kleiberg, Prabhakar Raghavan: Clustering categorical data: an approach based on dynamic systems". Proc. 1998 Int. Conf. On Very Large Databases, pp. 311-323, New York, August 1998.
[13] J.C. Gower, A general coefficient of similarity and some of its properties. BioMetrics, 27: 857-874, 1971.
[14] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, ROCK: A robust clustering algorithm for categorical attributes. ICDE 1999: 512-521.
[15] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. In Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98), pages 58-65, New York, NY, Aug. 1998.
[16] J. Han and M. Kamber, Data mining: concepts and techniques, Morgan Kaufmann publishers, 2001.
[17] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery, vol. 2, no. 3, pp 283-304, 1998.
[18] A.K. Jain and R.C. Dubes, Algorithms for clustering data, Rentice Hall, 1988.
[19] L. Kaufman and P.J. Rousseeuw, Finding groups in data – An Introduction to Cluster Analysis in Knowledge, 1990.
[20] Lioyd. Learning square quantization in PCM. (published in IEEE Trans. Information Theory), 28:128-137, 1982), Technical Report, Bell Labs, 1957.
[21] Tao Li, Sheng Ma, Mitsunori Ogihara, Entropy-based criterion in categorical clustering. In Proceedings of The 2004, IEEE International Conference on Machine Learning (ICML 2004), pages 536-543.
[22] J. MacQueen. Some methods for classi¯cation and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Statist, Prob., 1:281-297, 1967.
[23] R.S. Michalski and R.E. Stephen, Automated construction of classification: conceptual clustering versus numerical taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 5(4): 396-410, 1983.
[24] J.R. Quinlan, Induction of decision trees, *Machine Learning,* vol. 1, no. 1, pp. 81-106, 1986.
[25] J.R. Quinlan, C4.5: Programs for machine learning. Morgan Kaufmann, 1993.
[26] H. Ralambondrainy, A conceptual version of the k-means algorithm. Pattern Recognition Letters, 16:1147-1157, 1995.
[27] Claude. E. Shannon, A mathematical theory of communication, Bell System Technical Journal, vol.27, pp. 379-423 and 623-656, July and October, 1948.
[28] M. Steinbach, G. Karypis, and V. Kumar, A comparison of document clustering techniques, In KDD workshop on Text Mining, 2000.
[29] L. Talavera and J. Béjar, Intergrating declarative knowledge in hierarchical clustering tasks. Proceedings of the International Symposium on Intelligent Data Analysis, pp. 211-222, Amsterdam, The Netherlands: Springer-Verlag, 1999.
[30] Y. Zhang, A. Fu, C. Cai, and P. Heng, Clustering categorical data, In Proc. 2000 IEEE Int. Conf. Data Engineering, San Deigo, USA, March 2000.

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.