Discretization of Continuous Valued Dimensions in OLAP Data Cubes

Sellappan Palaniappan

Department of Information Technology,
Malaysia University of Science and Technology,
47301 Petaling Jaya, Selangor, Malaysia

Tan Kim Hong

Department of Information Technology
Malaysia University of Science and Technology,
47301 Petaling Jaya, Selangor, Malaysia

Abstract

Continuous valued dimensions in OLAP data cubes are usually grouped into countable disjoint intervals using naïve methods such as equal width binning, histogram analysis, or splitting into intervals defined by domain experts according to their understanding of the data. This paper explores an integration of 'intelligent' discretization techniques currently available in data mining research into the construction of a SEER breast cancer survivability data cube with continuous dimension. Observational and empirical evaluations on the resulting cube with discretized intervals show that 'intelligent' discretization methods provide the same benefits to OLAP data cubes as in data mining algorithms, that is, they are able to simplify the data representation with minimal or no loss of information. Additionally, it was found that an unsupervised discretization method using *k*-means algorithm had exhibited equivalent performance as the supervised counterparts, namely, the entropy-based (ID3) and χ^2–based (CHAID) methods.

Key words:
OLAP, data mining, discretization, entropy, ID3, CHAID, k-means

1. Introduction

Data mining and On-Line Analytical Processing (OLAP) are two key technologies highly discussed in data warehousing architecture. Initially, both technologies were focusing on respective non-overlapping functionalities to support warehouse data analysis i.e. OLAP technology was concentrated on enhancing the interaction and visualization of data, but lacking the functionalities to guide user on the drill-path to locate interesting information, whereas data mining applications automated discovery of implicit patterns and interesting knowledge hidden in the large amount of data, but they do not facilitate user-friendly exploration interface to present the mined result [1].

Knowing that both OLAP and data mining can complement each other to better support user's data analysis, many efforts from the research community have started to gear towards integrating them as a combined component in data warehousing implementations. [1] has introduced a new On-Line Analytical Mining (OLAM) server in his commercially available software DBMiner™ 2.0. [4] initiated the i^3 (i-cube) project on discovery-driven exploration of OLAP data cubes, and [6], [7] designed an advanced cube operator called cubegrade and a constrained gradient analysis that can calculate and highlight "grade-of-change" between surrounding cube cells using association mining rules.

In this paper, we are looking into another area of integration between OLAP and data mining, that is, the discretization of continuous-valued data. Discretization is a process that generalizes an attribute from the ratio or interval scale that may contain infinitely many data values into a set of countable disjoint intervals and thereby reduces and simplifies the original data. Currently, 'intelligent' discretization methods are deployed as either built-in functions or implemented as data pre-processing steps in many data mining algorithms such as ID3, C4.5, CART, CHAID, Naïve Bayesian classifier and association mining. Generally, mined results using discretized values are more usable, easier to understand and closer to human knowledge-level representation. However, OLAP tools still do not incorporate discretization as an automated function in their applications. Continuous valued dimensions are usually split into intervals manually using naïve methods such as binning and histogram analysis, or based on the knowledge of the data by domain experts without making inference from the database. Therefore, we are keen to find out if these 'intelligent' discretization methods can be used in OLAP data cubes to produce cubes that are simplified, but still best preserve the original underlying data distribution and relationships among variables.

1.1 Problem Statement

This paper explores an integration of both supervised and unsupervised discretization methods available in data mining applications into construction of an OLAP data cube with continuous valued dimension. The integrated model must be proven to solve the following problem statements:

1. How to identify an optimal number of intervals to discretize the data? Can the 'intelligent' discretization methods from data mining techniques automate this decision?
2. How do the 'intelligent' discretization methods enhance the discretization result as compared to naïve

methods used in existing OLAP data cube construction?

3. Do supervised discretization methods outperform unsupervised ones? Is class information used in supervised discretization methods mandatory?

In order to answer those questions, our goal of analysis is three fold: simplicity, consistency, and accuracy:

- Simplicity – discretized intervals simplify the data views in the cube
- Consistency – discretized intervals preserve original data distribution
- Accuracy – discretized intervals retain relationships among the variables

The rest of this paper is organized as follows. In the next section, we discuss the discretization process and a few of the widely used discretization techniques in data mining. Section 3 explains the SEER breast cancer data set used in our experiment, follows by detailed experiment framework design in section 4. Section 5 presents findings by performing an empirical evaluation on the experiment results. Finally the conclusions are given in section 6, with suggestions of future works related to this paper.

2. Discretization

Data discretization is a general purpose pre-processing method that reduces the number of distinct values for a given continuous variable by dividing its range into a finite set of disjoint intervals, and then relates these intervals with meaningful labels. Subsequently, data are analyzed or reported at this higher level of knowledge representation rather than the subtle individual values, and thus leads to a simplified data representation in data exploration and data mining process. [12] has formally defined a discretization process flow in four steps as depicted in Fig. 1: (1) sorting the continuous values of the attribute to be discretized, (2) evaluating a cut-point for splitting or adjacent intervals for merging, (3) according to some criterion, splitting or merging intervals of continuous values, and (4) finally stopping at some point based on a stopping criteria.

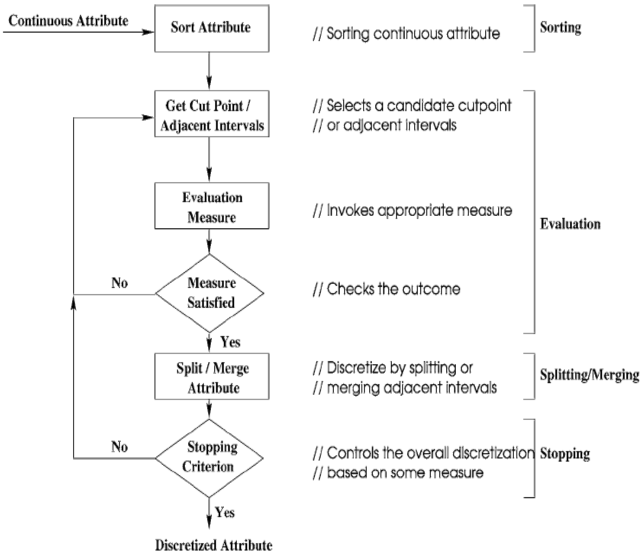


Fig. 1: Discretization process (Liu H. *et al.*, 2002)

To achieve simplified data representation using the different discretization techniques, we must not sacrifice information. Careful selection of a discretization technique that minimizes the loss of information is needed before hand. [11,12,19,20] have tried to make comparison of discretization methods using improved techniques to prove this requirement. [23] pointed out that effective discretization methods chosen can even produce new and more accurate knowledge. In following sections, we discuss in detail a few of the commonly used discretization techniques.

2.1 Equal Width & Equal Frequency Binning

The equal width binning is the simplest unsupervised method. The algorithm first sort the continuous valued attribute, then find the minimum x_{min} and the maximum x_{max} of that attribute. Interval width, w, is then computed by

$$w = \frac{x_{max} - x_{min}}{k}$$

where k is a user-specified parameter stating the number of intervals to discretize. The interval boundaries are specified as $x_{min} + w_i$, where $i = 1,2, \ldots, k-1$.

Equal frequency binning divides the sorted continuous values into k intervals such that each interval contains approximately ${}^{n}/_{k}$ data instances with adjacent values. Note that data instances with identical value must be placed in the same interval, thus it is not always possible to generate exactly k equal frequency intervals.

2.2 Interval Merging and Splitting using χ^2 Analysis

χ^2 is a statistic used to perform statistical independence test on relationship between two variables in a contingency table. In a database with data instances labeled with *p* classes, the formula to compute χ^2 statistic at a split point for two adjacent intervals against the *p* class values is:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{p} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

where:

p = number of classes,
A_{ij} = number of distinct values in the ith interval, jth class,
R_i = number of examples in ith interval $= \sum_{j=1}^{p} A_{ij}$,
C_j = number of examples in jth class $= \sum_{i=1}^{m} A_{ij}$,
N = total number of examples $= \sum_{j=1}^{p} C_j$ and
E_{ij} = expected frequency of $A_{ij} = (R_i \times C_j)/N$.

The test statistic has a χ^2 distribution or approximate to a χ^2 distribution if sample size is large enough and it measures the departure from H_0 (null) hypothesis which states that the two variables are statistically independent. The aim of χ^2 discretization is to merge adjacent intervals based on χ^2 statistic to obtain the most similar distribution with discretized intervals to the distribution of original data against the class values. Since the class frequencies are used in the computation of statistic, it is thus classified as a supervised discretization technique.

The popular CHAID (chi-squared automatic interaction detection) proposed by Kass (1980) is a top-down discretization algorithm that uses χ^2 statistic. It starts with one interval for the whole range, based on the *p*-values from χ^2 distribution; it determines the best next split at each step to further split the intervals. CHAID algorithm is being used as a splitting criterion in decision tree induction of many data mining software.

2.3 Entropy-based Discretization

In information theory, the entropy function for a given set *S*, or the expected information needed to classify a data instance in *S*, *Info(S)* is calculated as

$Info(S) = -\Sigma\, p_i \log_2 (p_i)$

where p_i is the probability of class *i* and is estimated as C_i/S, C_i being the total number of data instances that are of class *i*. A log function to the base *2* is used because the information is encoded in bits. The entropy value is bounded from below by *0*, when the model has no uncertainty at all, i.e. all data instances in *S* belong to one of the class $p_i = 1$, and other classes contain 0 instances $p_j = 0,\ i \neq j$. And it is bounded from the top by $\log_2 m$, where *m* is the number of classes in *S*, i.e. data instances are uniformly distributed across *k* classes such that $p_i = {}^1/_m$ for all.

Based on this entropy measure, J. Ross Quinlan (1986) developed an algorithm called Iterative Dichotomiser 3 D3) to induce best split point in decision trees. ID3 nploys a greedy search to find potential split-points within e existing range of continuous values using the following rmula:

$$\boldsymbol{Info(S,T)} = -p_{left} \sum_{j=1}^{m} p_{j,left} \log p_{j,left} - p_{right} \sum_{j=1}^{m} p_{j,right} \log p_{j,right}.$$

the equation, $p_{j,left}$ and $p_{j,right}$ are probabilities that an stances, belong to class *j,* is on the left or right side of a tential split-point *T*. The split-point with the lowest tropy is chosen to split the range into two intervals, and the binary split is continued with each part until a stopping criterion is satisfied. [20] propose a stopping criterion for this generalization using the minimum description length principle (MDLP) that stops the splitting when

$InfoGain(S,T) = Info(S) - Info(S,T) < \delta$

where *T* is a potential interval boundary that splits *S* into S_1 (left) and S_2 (right) parts, and

$$\delta = [\log_2(n-1) + \log_2(3^k - 2) - [m\, Info(S) - m_1\, Info(S_1) - m_2\, Info(S_2)]] / n$$

where m_i is the number of classes in each set S_i and *n* is the total number of data instances in *S*.

2.4 Discretization using Clustering Analysis

The most popular algorithm in clustering analysis, *k*-means by MacQueen (1967) is also suitable to be used to discretize continuous valued variables because it calculates continuous distance-based similarity measure to cluster data points. In fact, since unsupervised discretization involves only one variable, it is equivalent to a "1-dimensional" *k*-means clustering analysis.

k-means is a non-hierarchical partitioning clustering algorithm that, initially, a set of points called cluster seeds is selected as a first guess as the means of the clusters, then remaining data points are assigned to their respective nearest cluster seed to form temporary clusters, after that seeds are replaced by the means of the temporary clusters, and this process is repeated until an optimum least-squares criterion is found or convergence is achieved i.e. no further changes occur in the clusters. Theoretically, clusters formed in this way should minimize the sum of squared distance between data points within each cluster over the sum of squared distance between data points from different clusters. The most common distance measure used in *k*-means algorithm is the Euclidean distance, a special case (*p=2*) of the Minkowski metric:

$$d_p(\mathbf{x}_i, \mathbf{x}_j) = (\sum_{k=1}^{d} |x_{i,k} - x_{j,k}|^p)^{1/p}$$
$$= \|\mathbf{x}_i - \mathbf{x}_j\|_p.$$

Other clustering algorithms were also tried as discretization methods such as the Proportional k-interval discretization (PKID) by [13]. [19] from Microsoft Research group has improved a mixture model clustering algorithm called Expectation Maximization that assigns each data point to each cluster according to a weight representing its probability of membership and has integrated it into Microsoft SQL Server™ Analysis Services (SSAS) applications as a discretization method to automatically discretize continuous dimension into buckets during OLAP data cube construction and data mining modeling.

3. SEER Breast Cancer Data Set

In medical decision support, many data perspectives from disease attributes and health measures are of numeric type and continuously valued. For example, the diagnostic characteristics and pathology measures of breast cancer patients such as tumor size, clump thickness, hormone receptors percentage, lymphatic invasion means and so on are mostly continuously-valued. Cut-points estimation to transform these continuous measures into groups of values that reflect the biological threshold effect is not as trivial as data from other industries. The choice of cut-points generally derived from either biological knowledge about the particular prognostic risk factor or physician's experience or the results already published in other studies. For some newly identified or previously unexplored prognostic factors, statistical methods mostly derived from classical regression theories such as log rank or Mantel-Cox test, likelihood ratio test and Wald statistics are used to estimate optimal cut-points on these continuous variables. Newer discretization techniques invented in data mining studies are rarely tried in medical analysis as methods to categorize continuous variables. Therefore we are of the opinion that a medical data set will serve as a good candidate to be used in our study to test and verify these newer discretization methods.

The Surveillance, Epidemiology, and End Results (SEER) Program that is managed by National Cancer Institute (NCI) is an authoritative source of information on cancer incidence and survival in the United States. It collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 26 percent of the US population. The SEER Program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, extent of disease (EOD), first course of treatment, and follow-up for vital status. The SEER cancer data set facilitates all kind of analysis dealing with cancer prevention, mortality, extent of disease at diagnosis, therapy and patient survival.

Among the cancer data provided (Note: A signed Limited-Use Data Agreement is required to access SEER data at URL http://seer.cancer.gov/data/access.html), the breast cancer data set from "SEER 1973-2005 Limited-Use Data" has been chosen. In absence of a medical domain expert involvement, it is not an easy task to meaningfully pick and extract the relevant attributes from the raw cancer data file. Since the emphasis of our study is from the angle of how different discretization techniques can help to simplify data representation in OLAP data cubes exploration with minimal loss of information, we have decided to make reference to work done by [22] for this purpose.

4. Experiment Framework

SEER breast cancer data is obtained in raw ASCII text file. Essentially, steps needed to process the raw SEER breast cancer data file into the final SEER breast cancer survivability OLAP data cube are designed as in the process flow diagram shown in Fig. 2.

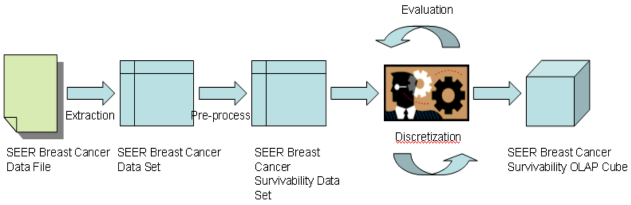

Evaluation
Extraction
Pre-process
Discretization
SEER Breast Cancer Data File
SEER Breast Cancer Data Set
SEER Breast Cancer Survivability Data Set
SEER Breast Cancer Survivability OLAP Cube

Fig. 2: Process Flow to Construct SEER Breast Cancer Survivability Cube

These steps can be further explained as follows:

- Extraction – to extract and covert all data items of all breast cancer incidences and store them in a structured SAS data set for pre-processing.
- Pre-process – to filter incomplete data records, to subset only selected data items relevant to survival analysis, and to transform and derive additional data items based on existing data items then label the

survival status for each patience incidence. A preliminary analysis on the resulting SEER breast cancer survivability data set is to be performed to confirm the sample created is valid and meaningful.

- Discretization – based on subject of analysis, identify measures and their calculation, and the dimensions from corresponding attributes for cube construction. Usually, dimension attributes are of categorical type, some numeric attributes may also be used as dimensions. As for a dimension attribute that is continuously valued, discretize and transform them to a reduced countable number of disjoint intervals to simplify the data representation. This part of work serves as the core of our analysis. A continuous dimension will be derived and discretized with a series of selected discretization techniques to produce a set of discretized intervals, subsequently empirically analyzed to determine optimum number of discretized intervals and method used.
- Cube construction – to design and build the SEER breast cancer survivability cube with the discretized intervals resulting from the previous steps.
- Cube exploration – to explore the cube with discretized intervals using SAS Enterprise Guide for survivability analysis. Observe how discretized intervals improve and simplify user's exploration experience.
- Empirical evaluation – an empirical evaluation of discretized intervals will then be conducted to measure their performance against the research objectives stated in the problem statements.

4.1 Extracting SEER Breast Cancer Text File

The SEER breast cancer text file contains 553,483 breast cancer cases diagnosed in the years of 1973-2005. These cases are extracted and stored into a structured data set containing 553,483 observations and 115 variables. This data set is then sorted in ascending order of "patient_id" so that records for the same patient are clustered together and is ready for pre-processing.

4.2 Pre-Processing SEER Breast Cancer Data Set

With reference to [22], the extracted SEER breast cancer data set is pre-processed by performing four major tasks: (1) to exclude incomplete cases especially those with unknown information, (2) standardization of "Site specific surgery code", (3) derive class labels, and (4) format coded values in raw extraction to descriptive values based on a series of lookup mappings. The final data set contains 114,142 observations and the 16 variables as selected by [22] shown in Fig. 3, plus an additional variable "survive" to store the class label.

Nominal variable name	Number of distinct values
Race	27
Marital status	5
Primary site code	9
Histologic type	20
Behavior code	2
Grade	5
Extension of tumor	29
Lymph node involvement	9
Site specific surgery code	10
Radiation	9
Stage of cancer	4

Numeric variable name	Mean	Std Dev	Range
Age	57.4908798	12.9058361	11-106
Tumor size	20.5131678	19.1865843	0-919
No of positive nodes	1.4558708	3.7545406	0-75
Number of nodes	14.477125	7.0825773	1-90
Number of primaries	1.1810289	0.4441687	1-8

Fig. 3: Characterization of SEER Breast Cancer Survivability Data Set Attributes

For the class distribution, about 86.7% or 98,997 out of 114,142 cases have been identified as "survived" and the remaining 13.3% or 15,145 out of 114,142 cases are classified as "not survived" as shown in Fig. 4.

Class	No of instances	Percentage
0: not survived	15,145	13.3
1: survived	98,997	86.7
Total	114,142	100.0

Fig. 4: SEER Breast Cancer Survivability Class Distribution

4.3 Identification of Continuous Dimension

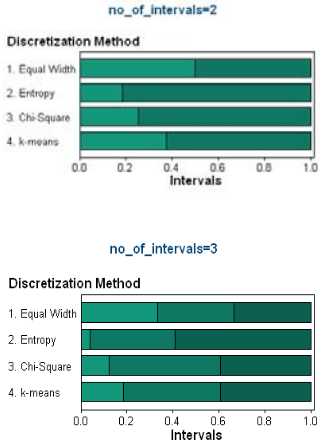
Existing clinical researchers relate "stage of cancer" to be the most relevant attribute in breast cancer survival analysis and the most common system used to describe the stages of breast cancer is the AJCC/TNM system defined by the American Joint Committee on Cancer. AJCC/TNM system takes into account the tumor size and spread (T), whether the cancer has spread to lymph nodes (N), and whether it has spread to distant organs (M, for metastasis). In our pre-processed SEER breast cancer data set, relative value of two variables, which is "no of positive nodes"/ "number of nodes", seems to express the factor N: spread to lymph nodes defined in the AJCC/TNM system better than if they are being used separately.

This is a continuous dimension. Even if it is rounded to a numeric value of two decimal places, there are still 100 possible distinct values. Higher precision will cause the number of categories to multiply, making the generated cube having very sparse (in which most of the cells are zero) cuboids on this dimension. Therefore, this continuous dimension, which is named "percent_of_positive_nodes" will be discretized into *k* intervals with a series of discretization methods such that the most suitable discretization method and an optimum number of intervals can be identified to simplify its data representation, but still able to preserve the original data distribution as well as retaining the variables relationships as much as possible.

4.4 Discretization of Continuous Dimension

In finding the optimal discretization method, four discretization techniques are tested on the continuous dimension "percent_of_positive_nodes": equal width, entropy-based (ID3), χ^2-based (CHAID) and *k*-means clustering.

Discretized intervals by each technique are outlined graphically using stacked bars as illustrated in Fig. 5. It is easily perceived that 2-interval and 3-interval are not the desirable number of intervals, as the split-points differ substantially from one method to the others. Entropy-based (ID3) and χ^2–based (CHAID) techniques quickly resemble each other starting at 4-interval splitting and producing very similar split-points for 5-, 6- and 7-interval splittings. This corresponds exactly to the findings by [23] stating that "*there is a somewhat unexpected connection between discretization methods based on information theoretical complexity, on one hand, and the methods which are based on statistical measures of the data dependency of the contingency table, such as Pearson's χ^2 or G^2 statistics on the other hand.*"



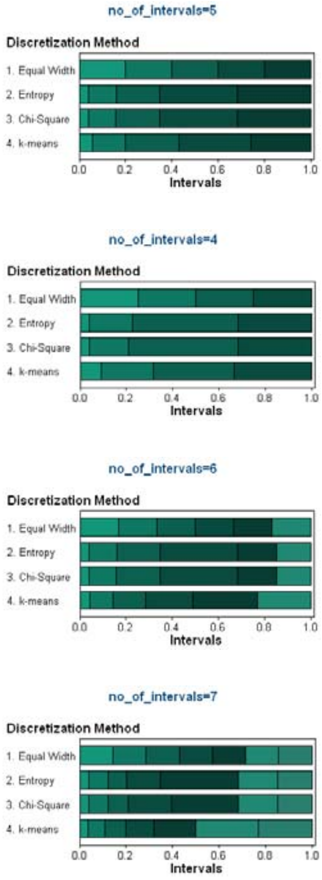


Fig. 5: 2-interval to 7-interval discretization by each of four discretization methods tested

Surprisingly, the split-points generated by *k*-means clustering have shown tendency to liken those of supervised methods when number of intervals increases. This phenomenon definitely poses a challenge to the common sayings found in discretization literature stating that supervised methods will always outperform or more superior to those methods that are not supervised with a class label [11]. The *k*-means algorithm is a minimum square error partitioning method that generates an arbitrary number *k* of partitions reflecting the original distribution of the partition attribute (Duda and Hart, 1972). This characteristic agrees well with another emphasis (i.e. preserving the original data distribution) in supervised algorithms apart from focusing on class probabilities and may help to explain the above observed coincidence.

Lastly, with so many disparities exist in discretized intervals by equal width method comparing to those by 'intelligent' methods, it is evident that naïve methods such as equal width which does not make statistical inference from the underlying data should be fallen into disuse.

5. Empirical Evaluation

Criteria of testing will focus on the three parameters that we intended to achieve in our research objectives:

- Simplicity – discretized intervals simplify the data views in the cube
- Consistency – discretized intervals preserve original data distribution
- Accuracy – discretized intervals retain relationships among the variables

5.1 Simplicity

A straightforward approach to determine simplicity is the reduction in cell counts with discretized intervals. Undiscretized “percent_of_positive_nodes” in SEER breast cancer data set has 534 distinct values (stored in IBM double-wide 8-byte floating point format with full precision by the SAS software), a 2-interval discretization will reduce the cell counts by 532 for each crossing with this dimension in a cube view and able to shrink it with a factor of 532/534=99.63%, 3-interval discretization will reduce it with a factor of 99.44% and so on as calculated in Table 1.

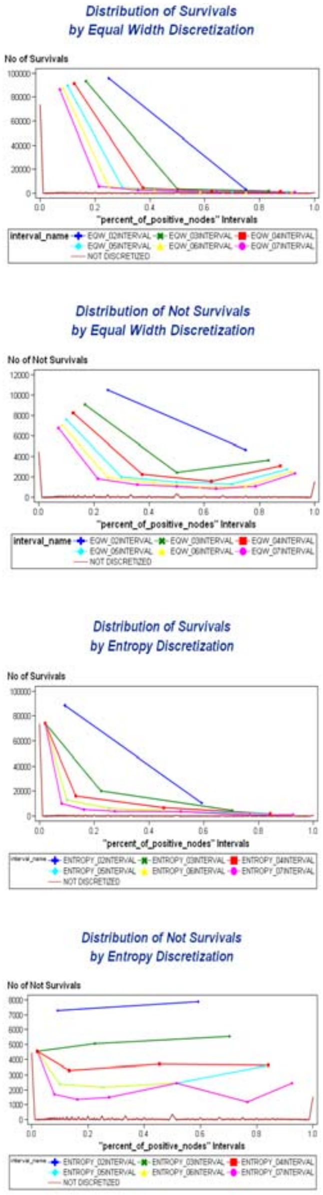
No of Intervals	% of reduction in cell counts
2	99.63%
3	99.44%
4	99.25%
5	99.06%
6	98.88%
7	98.69%
8	98.50%
9	98.31%
10	98.13%
11	97.94%
12	97.75%
13	97.57%
14	97.38%
15	97.19%

Table 1: % of reduction in cell counts with discretized intervals

The 7-inteval discretization has been tested optimal in our experiment hence we can conclude that the SEER breast cancer survivability cube constructed obtains simplicity of 98.69% for each crossing with this discretized dimension in cube views.

5.2 Consistency

To evaluate how discretized intervals preserve the original data distribution, Fig. 6 charted the distribution plots for “survive” and “not survive” class label with different number of intervals by each of four methods overlaid with the original data distribution to illustrate the degree of resemblance or difference for each discretization technique experimented.

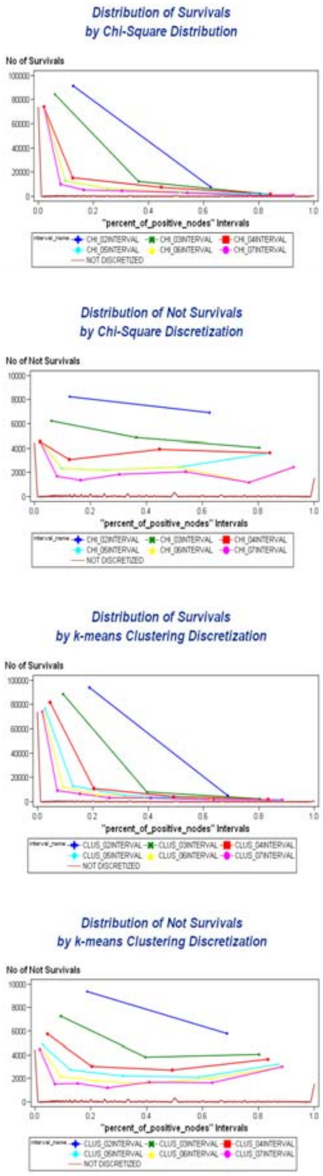


Distribution of Survivals by Equal Width Discretization

Distribution of Not Survivals by Equal Width Discretization

Distribution of Survivals by Entropy Discretization

Distribution of Not Survivals by Entropy Discretization



Distribution of Survivals by Chi-Square Distribution

Distribution of Not Survivals by Chi-Square Discretization

Distribution of Survivals by k-means Clustering Discretization

Distribution of Not Survivals by k-means Clustering Discretization

Fig. 6: Distribution of "survive" and "not survive" versus discretized intervals by each technique

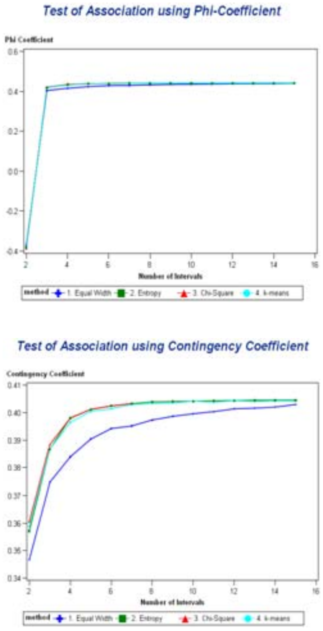
A quick run through of the distribution plots reveals the intervals discretized with naïve equal width methods deviate from the original data distribution the most, the entropy-based (ID3) and χ^2-based (CHAID) discretizations produce similar intervals and they start exhibiting resemblance to the original data distribution at 4-interval splitting. *k*-means clustered intervals converge as good as the supervised counter-parts, the entropy-based (ID3) and χ^2–based (CHAID) discretization methods, except that they converge slower. Somehow this observation answers the 3rd question in our problem statement, suggesting that unsupervised methods like *k*-means clustering can perform equally to that of supervised methods simply due to its characteristic of being an algorithm that uses minimum square error partitioning to generate an arbitrary number *k* of partitions reflecting the original distribution of the partition attribute (Duda and Hart, 1972).

5.3 Accuracy

We shall adopt the idea of measuring the accuracy criterion on how discretized intervals retain relationships among the variables using statistical tests for contingency tables. Two widely used tests on statistical significance of the variable relationships in contingency tables are test of association and test of independence.

Test of Association

The degree of association between two variables can be assessed by a number of coefficients, the simplest are the *phi* and *contingency* coefficients. Both *phi* and *contingency* coefficients are calculated for contingency tables of discretized variables versus class label "survive" for the four discretization techniques experimented: equal width, entropy-based (ID3), χ^2-based (CHAID) and *k*-means clustering, Fig. 7 plots their line charts of both coefficients calculated for 2-interval to 15-interval discretization.



Test of Association using Phi-Coefficient

Test of Association using Contingency Coefficient

Fig. 7: Test of association for discretized variable vs class label

The undiscretized variable is associated with the class label "survive" with strength of *phi* value 0.4520 and *contingency* value 0.4119. From the line plots, we see this convergence starts at 4-interval discretization for all three 'intelligent' methods and approaching the desirable optimum from 7-

interval onwards. Slow convergence by the naïve method equal width shown in *contingency* plot again reassures our previous observations that naïve methods should not be used to discretize continuous valued dimension.

Test of Independence

Another way to treat discretization is to merge intervals so that the rows (intervals) and columns (classes) of the contingency table become more statistically dependent [23]. The Pearson's χ^2 and the likelihood-ratio statistic G^2 are used similarly to test the independence of the null hypothesis $H_{0.}$

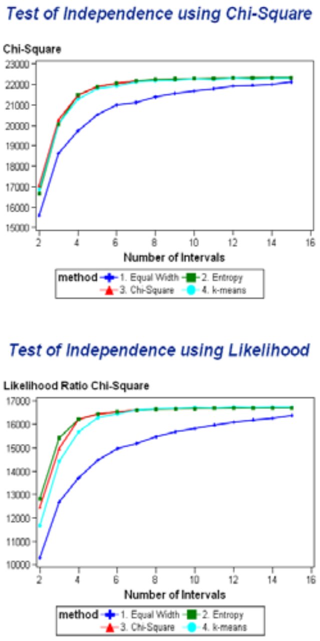


Fig. 8: Test of independence for discretized variable vs class label

Fig. 8 shows plots of Pearson's χ^2 and G^2 statistics computed. The convergence approaching toward the statistics values of undescretized data Pearson's χ^2=23322.5238 and G^2=17365.9454 is readily seen. These high values clearly indicate strong departure from null hypothesis H_0, and thus failed the assumption that the discretized variable is independent of "survive" class label. As usual, naïve method still performs the worst.

Both test of association and test of independence conducted obviously signify that discretized intervals are able to preserve original variable relationships, as long as an optimal number of intervals can be identified.

6. Conclusion and Future Work

We summarize this research by answering the stated problem statements by making reference to observations and test results gathered in the experiment as follows:

1. How to identify an optimal number of intervals to discretize the data? Can the 'intelligent' discretization methods from data mining techniques automate this decision?

   In our experiment, various tests and observations have confirmed that the optimal number of intervals is 7. This number was first identified by the automatic pruning criterion of statistical significance test using p-value when we discretized the "percent_of_positive_nodes" using χ^2–based (CHAID) algorithm. This suggests that a statistical significance test using p-value can be deployed in our integrated model to automate the decision for number of intervals in discretization problems.

2. How do the 'intelligent' discretization methods enhance the discretization result as compared to naïve methods used in existing OLAP data cube construction?

   Many test results gathered in empirical evaluation, especially in the sections which measure the consistency and accuracy of discretized results indicate that the 'intelligent' discretization methods clearly out weigh the selection of naïve methods which do not make statistical inference from the database. By not doing so, naïve methods fail to preserve original data distribution and existence of variable relationships, rendering the data cube exploration ineffective.

3. Do supervised discretization methods outperform unsupervised ones? Is class information used in supervised discretization methods mandatory?

   This finding is interesting. As we have observed in Section 4.4 and 5, k-means clustering generates split-points likened to those of supervised methods when number of intervals increase. This phenomenon definitely poses a challenge to the common saying found in most discretization literature that supervised methods will always outperform or superior to those methods that are not supervised with a class label [11]. Somehow this observation also suggests that unsupervised methods like k-means clustering can perform equally well to that of supervised methods simply due to its characteristic of being an algorithm that uses minimum square error partitioning to generate an arbitrary number k of partitions reflecting the original distribution of the partition attribute (Duda *and Hart, 1972), and thus class information is not mandatory in discretization problems.*

The integrated model combining data mining functionalities into OLAP applications can be enormous, but our focus is in simplifying data representation of continuous dimensions frequently found in OLAP data cubes using discretization techniques available from data mining literature with the aim of minimizing information loss in cube exploration. Other techniques in data mining studies that are appropriate to the purpose of this problem include the following:

1. Multivariate discretization – Many discretization algorithms developed in data mining field focus on univariate, which discretize each continuous attribute independently, without considering interactions with other attributes, at most taking the interdependent relationship between class attribute into account likes what we saw in supervised discretization techniques. Due to the nature of multi-dimensional space of OLAP construction, discretizing continuous dimension by taking care of a single class label is not sufficient. Relationships with other non-class dimensions are equivalently important. Works done in this area include multivariate discretization methods for set mining proposed by Bay (2001), multivariate interdependent discretization using Bayesian network structure by S. Monti *et al*. (1998), and others.
2. Multilevel discretization – Discretization can also be performed recursively on an attribute to provide a hierarchical or multiresolution partitioning of the attributes values, known as concept hierarchy that are useful for mining at multiple levels of abstraction. This capability is definitely well suited for OLAP data cube as concept hierarchy is common in cube dimensions.

References

[1] J. Han and M. Kambler (2006) *Data Mining: Concepts and Techniques* (2nd edition). Morgan Kaufmann, San Francisco.

[2] S. Chaudhuri and U. Dayal. (1997) An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, March 1997

[3] S. Sarawagi, Rakesh Agrawal and Nimrod Megiddo (1998) Discovery-driven exploration of OLAP data cube. *Proc. of the Sixth Int'l Conference on Extending Database Technology (EDBT)*, Valencia, Spain, Page 168-182, March 1998.

[4] S. Sarawagi, G. Sathe. (2000) i^3: Intelligent, Interactive Investigation of OLAP data cubes. ACM SIGMOD Record, Volume 29 Issue 2, June 2000.

[5] J. Han (1998) Towards On-Line Analytical Mining in Large Databases. ACM SIGMOD Record, 27:97-107, March 1998.

[6] T. Imielinski, L. Khachiyan and A. Abdulghani. (2002) Cubegrades: Generalizing Association Rules. *Data Mining and Knowledge Discovery,* 6(3):219-258, 2002.

[7] G. Dong, J. Han, J.M.W. Lam *et al.* (2001) Mining Multi-Dimensional Constrained Gradients in Data Cubes. *Proc. of the 27th Int'l Conference on Very Large Data Bases (VLDB)*, Page 321-330, 2001.

[8] S. Sarawagi (2000) User-adaptive exploration of multidimensional data. *Proc. of the 26th VLDB Conference,* Cairo, Egypt, 2000.

[9] S. Sarawagi (2001) User-cognizant multidimensional analysis. *The VLDB Journal,* Volume 10 Issue 2-3, Page 224-239. Springer-Verlag New York, Inc, September 2001.

[10] E.F. Codd Associates (1998) *Providing OLAP to User-Analysts: An IT Mandate*. White paper available from http://www.uniriotec.br/~tanaka/SAIN/providing_olap_to_user_analysts.pdf

[11] J. Dougherty, R. Kohavi, M. Shami (1995) Supervised and Unsupervised Discretization of Continuous Features. *Proc. of the 12th Int'l Conference on Machine Learning*, Page 194-202, 1995

[12] H. Liu, F. Hussain, C.L. Tan, & M. Dash (2002) Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery,* 6:393-423, 2002

[13] Ying Yang, Geoffrey I. Webb (2001) Proportional k-Interval Discretization for Naïve-Bayes Classifiers, *Proc. of the 12th European Conference on Machine Learning* (ECML01) pp 564-575

[14] Ying Yang, Geoffrey I. Webb (2003) *Discretization for naïve-Bayes learning: managing discretization bias and variance.* Technical Report 2003/131 School of Computer Science & Software Engineering, Monash University

[15] Ying Yang, Geoffrey I. Webb (2003) O*n Why Discretization Works for Naïve-Bayes Classifiers*. School of Computer Science and Software Engineering, Monash University

[16] Michael K. Ismail, Vic Ciesielski (2003), An Empirical Investigation of the Impact of Discretization on Common Data Distributions. *Proc. of the 3rd Int'l Conference on Hybrid Intelligent System (HIS'03)*, Page 692-701, December 2003

[17] Chotirat Ann R. (2003) CloNI: Clustering of $\sqrt{N}$ -Interval discretization. *Proc. of 4th Int'l Conference on Data Mining*, December 2003

[18] Marco Vannucci, Valentina Colla (2004) Meaningful discretization of continuous features for association rules mining by means of a SOM. *European Symposium on Artificial Neural Networks 2004* pp 489-494, April 2004

[19] Bradley, Fayyad, and Reina (1998) *Scaling EM (Expectation-Maximization) Clustering to Large Databases.* Technical Report MSR-TR-98-35, Microsoft Research

[20] Fayyad U.M., Irani K.B. (1993) Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *Proc. of 13th Int'l Joint Conference on Artificial Intelligence.* pp. 1022-1027

[21] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Limited-Use Data (1973-2005), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2008, based on the November 2007 submission.

[22] Abdelghani Bellaachia, Erhan Guven (2006) Predicting Breast Cancer Survivability Using Data Mining Techniques. Department of Computer Science, George Washington University.

[23] Ruoming J., Yuri B. (2007) Data Discretization Unification. Department of Computer Science, Kent State University, March 2007

Sellappan Palaniappan obtained his PhD in Interdisciplinary Information Science from University of Pittsburgh and a MSc in Computer Science from University of London. He is an Associate Professor at the Department of Information Technology, Malaysia University of Science and Technology. His research interests include information integration, clinical decision support systems, OLAP and data mining, web services and collaborative CASE tools.

Tan Kim Hong obtained her degree in Mathematics from University of Malaya. She is a freelance IT consultant in data warehousing and business intelligence. Currently, she is doing her MSc degree in Information Technology at Malaysia University of Science and Technology (MUST).