Automatic KDD Data Preparation Using Multi-criteria Features

Youssef Hmamouche*, Christian Ernst† and Alain Casali*
*LIF - CNRS UMR 6166, Aix Marseille Université, Marseille, France
Email: firstname.lastname@lif.univ-mrs.fr
†CMP - SGC, Ecole des Mines de St Etienne and LIMOS, CNRS UMR 6158, Gardanne, France
Email: ernst@emse.fr

***Abstract*—We present a new approach for automatic data preparation, applicable in most Knowledge Discovery and Data Mining systems, and using statistical features of the studied database. First, we detect outliers using an approach based on whether data distribution is normal or not. We outline further that, when trying to find the most appropriate discretization method, what is important is not the law followed by a column, but the shape of its density function. That is why we propose an automatic choice for finding the best discretization method based on a multi-criteria (Entropy, Variance, Stability) analysis. Experimental evaluations validate our approach: The very same discretization method is never always the most appropriate.**

Keywords–Data Mining; Data Preparation; Outliers; Discretization Methods.

I. Introduction and Motivation

Data preparation can be performed according to different method(ologie)s [1]. However, this task has not been developed greatly in the literature: The single mining step is more often emphasized. Moreover, it focuses most of the times on a single parameter: discretization method [2], outlier detection [3], null values management, *etc.*. Associated proposals only highlight on their advantages comparing themselves to others. There is no global nor automatic approach taking advantage of all of them. But the better data are prepared, the better results will be, and the faster mining algorithms will work. Previously in [4], we proposed a simple but efficient approach to transform input data into a set of intervals (also called bins, clusters, classes, *etc.*). On which we apply, in a further step, specific mining algorithms (correlation rules, *etc.*). The reasons that decided us to reconsider previous works are: (i) To improve the outliers' detection with regard to the data distribution (normal or not), (ii) To reduce the number of input parameters, and thus (iii) To propose an automatic choice of the best discretization method. Finally, regarding implementation, we merge the three tasks into a single one, and carry out experiments.

This paper is organized as follows: Section II presents general aspects of data preparation. Section III and Section IV are dedicated to outlier detection and to discretization methods respectively. Each section is composed of two parts: (i) related work, and (ii) our approach for improving it. In Section V, we show the results of first experiments. Last Section summarizes our contribution, and outlines some research perspectives.

II. Data Preparation

Raw input data must be prepared in any Knowledge and Discovery in Databases (KDD) system previous to the mining step. There are two main reasons for this:

- If each value of each column is considered as a single item, there will be a combinatorial explosion of the search space, and thus very large reponse times;
- We cannot expect this task to be performed by hand because manual cleaning of data is time consuming and subject to many errors.

Generally, this step is divided into two tasks: (i) Preprocessing, and (ii) Transformation(s).

A. Preprocessing

Preprocessing consists in reducing the data structure by eliminating columns and rows of low significance [5].

a) Basic Column Elimination: Elimination of a column can be the result of, for example in the microelectronic industry, a sensor dysfunction, or the occurrence of a maintenance step; this implies that the sensor cannot transmit its values to the database. As a consequence, the associated column will contain many null/default values and must then be deleted from the input file. Elimination should be performed by using the Maximum Null Values ($MaxNV$) threshold. Furthermore, sometimes several sensors measure the same information, what produces identical columns in the database. In such a case, only a single column should be kept.

b) Elimination of Concentrated Data and Outliers: We first turn our attention to inconsistent values, such as "outliers" in noisy columns. Detection should be performed through another threshold (a convenient value of p when using the standardization method, see Section III-A1). Found outliers are eliminated by forcing their values to Null. Another technique is to eliminate the columns that have a small standard deviation (threshold $MinStd$): Since their values are almost the same, we can assume that they do not have a significant impact on results; but their presence pollutes the search space and reduces response times. Similarly, the number of Distinct Values in each column should be bounded by the minimum ($MinDV$) and the maximum ($MaxDV$) values allowed.

B. Transformation

c) Data Normalization: This step is optional. It translates numeric values into a set of values between 0 and 1. Standardizing data simplifies their classification.

d) Discretization: Discrete values deal with intervals of values, which are more concise to represent knowledge, so that they are easier to use and also more comprehensive than continuous values. Many discretization algorithms (*see* Section IV-A) have been proposed over the years for this. The number of used intervals ($NbBins$) as well as the selected discretization method among those available are here again parameters of the current step.

 ISBN: 978-1-61208-415-2

e) Pruning step: When the occurrence frequency of an interval is less than a given threshold ($MinSup$), then it is removed from the set of bins. If no bin remains in a column, then that column is entirely removed.

The presented thresholds/parameters are the ones we use for data preparation. In previous works, their values were fixed inside of a configuration file read by our software at setup. The objective of this work is to automatically determine most of these variables without information loss. Focus is set on outlier and discretization management.

III. Detecting Outliers

An outlier is an atypical or erroneous value corresponding to a false measurement, a calculation mistake, an unwritten input, *etc.* Outlier detection is an uncontrolled problem because extreme values deviate too greatly from the rest of the data. In other words, they are associated with a significant deviation from the other observations [3]. In this section, we present some outlier detection methods and focus on the detection of outliers in the case of uni-variate data.

The following notations are used to describe outliers: X is a numeric attribute of a database relation, and is increasingly ordered. x is an arbitrary value, X_i is the i^{th} value, N the size of X, σ its standard deviation, μ its mean, and s a central tendency parameter (variance, inter-quartile range, *etc.*). X_1 and X_n are the minimum and the maximum values of X respectively. p is an arbitrary probability, and k is a parameter specified by the user, or computed by the system.

A. Related Work

In this section, we summarize four of the principal univariate outlier detection methods.

1) Elimination after standardizing the distribution: This is the most conventional cleaning method [3]. It consists in taking into account μ and σ to determine the limits beyond which aberrant values will be eliminated. For an arbitrary distribution, the Bienaymé-Tchebyshev inequality specifies that the probability that the absolute deviation between a variable and its average is greater than p is less than or equal to $\frac{1}{p^2}$:

$$P(\left|\frac{x-\mu}{\sigma}\right| \geq p) \leq \frac{1}{p^2} \quad (1)$$

The idea is to set a threshold probability as a function of μ and σ above which we accept values as non-outliers. For example, with $p = 4.47$, the probability that x, satisfying $\left|\frac{x-\mu}{\sigma}\right| \geq p$, is an outlier is bounded by 0.05.

2) Algebraic method: This general detection method, presented in [6], uses the relative distance of a point to the "center" of the distribution: $d_i = \frac{|X_i-\mu|}{\sigma}$. Outliers are detected outside of the interval $[\mu - kQ_1, \mu + kQ_3]$, where k is generally set between 1.5 and 3. Q_1 and Q_3 are the first and the third quartiles respectively.

3) Box plot: This method, attributed to Tukey [7], does not make any assumption on how the data are distributed. It is based on the difference between quartiles Q_1 and Q_3, and distinguishes between two categories of extreme values determined outside the lower bound (LB) and the upper bound (UB):

$$\begin{cases} LB = Q_1 - k \times (Q_3 - Q_1) \\ UB = Q_3 + k \times (Q_3 - Q_1) \end{cases} \quad (2)$$

4) Grubbs' test: Grubbs' method, presented in [8], is a statistical test for lower or higher abnormal data. It uses the difference between the average and the extreme values of the sample. The test is based on the assumption that the data are normally distributed. The maximum and minimum values are tested, which allows one to decide if any of these values is aberrant. The statistic used is $T = max(\frac{X_N-\mu}{\sigma}, \frac{\mu-X_1}{\sigma})$. The test is based on two hypotheses:

- Hypothesis H_0: The tested value is not an outlier.
- Hypothesis H_1: The tested value is an outlier.

Hypothesis H_0 is rejected at significance level α if:

$$T > \frac{N-1}{\sqrt{n}} \sqrt{\frac{\beta}{n-2\beta}} \quad (3)$$

where $\beta = t_{\alpha/(2n),n-2}$ is the quartile of order $\alpha/(2n)$ of the Student distribution with $n-2$ degrees of freedom.

B. An Original Method for Outlier Detection

Many existing outlier detection methods assume that the distribution of data is normal. However, we observed that, in reality, many samples have asymmetric and/or multimodal distributions; the use of these methods will then have a significant influence on the mining step. Therefore, we should process each distribution using an appropriated method. The considered approach consists in eliminating outliers in each column based on the normality of data, in order to minimize the risk of eliminating normal values.

Firstly, the Kolmogorov-Smirnov test presented in [9] is applied in order to determine whether the distribution is normal or not. Secondly, if the variable is normally distributed, then the Grubbs' test is used at a significance level of 5%. This test gives experimentally better results than the algebraic approach. Otherwise, the *Box plot* method is employed with parameter k set to 3 in order to not to be too exhaustive toward outlier detection. Figure 1 summarizes the process we chose for detecting and deleting outliers.

In the previous versions of our software, we used the simple standardization method with p set as an input parameter. With this new approach, no input parameter remains. We obtained moreover an improvement of 2% in the detection of true positive or false negative outliers.

IV. Discretization Methods

Discretization of an attribute consists in finding $NbBins$ disjoint intervals that will further represent it in an efficient way. The final objective of discretization methods is to ensure that the mining part of the KDD process generates substantial results. In our approach, we only employ direct discretization methods in which $NbBins$ must be known in advance (and be the same for every column of the input data). $NbBins$ was initially a parameter fixed by the end-user. The literature proposes several formulas as an alternative (Rooks-Carruthers, Huntsberger, Scott, *etc.*) for computing such a

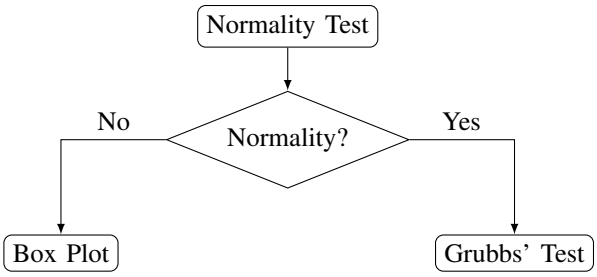

Figure 1. Main tests used in our outlier detection process.

number. Therefore, we switched to the Huntsberger formula, the most fitting from a theoretical point of view [10], and given by: $1 + 3.3 \times log_{10}(N)$.

A. Related Work

In this section, we only highlight the final discretization methods kept for this work. This is because the other tested methods have not revealed themselves to be as efficient as expected (such as Embedded Means Discretization), or are not a worthy alternative (such as Quantiles based Discretization) to the ones presented. The methods we use are: Equal Width Discretization (EWD), Equal Frequency-Jenks Discretization (EFD-Jenks), AVerage and STandard deviation based discretization (AVST), and K-Means (KMEANS). These methods, which are unsupervised [11] and static [12], have been widely discussed in the literature: See for example [10] for EWD and AVST, [13] for EFD-Jenks, or [2] and [14] for KMEANS. For these reasons, we only summarize their main characteristics and their field of applicability in Table I.

TABLE I. SUMMARY OF THE DISCRETIZATION METHODS USED.

Method	**Principle**	**Applicability**
EWD	This simple to implement method creates intervals of equal width.	The approach cannot be applied to asymmetric or multimodal distributions.
EFD-Jenks	Jenks' method provides classes with, if possible, the same number of values, while minimizing internal variance of intervals.	The method is effective from all statistical points of view but presents some complexity in the generation of the bins.
AVST	Bins are symmetrically centered around the mean and have a width equal to the standard deviation.	Intended only for normally distributed datasets.
KMEANS	Based on the Euclidean distance, this method determines a partition minimizing the quadratic error between the mean and the points of each interval.	The disadvantage of this method is its exponential complexity, so computation time can be long. It is applicable to each form of distribution.

Let us underline that the upper limit fixed to the number of intervals to use is not always reached, depending on the applied discretization method. Thus, EFD-Jenks and KMEANS generate most of the times less than $NbBins$ bins.

Example 1: Let us consider the numeric attribute $SX = \{4.04, 5.13, 5.93, 6.81, 7.42, 9.26, 15.34, 17.89, 19.42, 24.40, 25.46, 26.37\}$. SX contains 12 values, so by applying the Huntsberger's formula, if we aim to discretize this set, we have to use 4 bins.

Table II shows the bins obtained by applying all the discretization methods proposed in Table I. Table III shows the number of values of SX belonging to each bin associated to every discretization method.

TABLE II. SET OF BINS ASSOCIATED TO SAMPLE SX.

Method	Bin_1	Bin_2	Bin_3	Bin_4
EWD	[4.04, 9.62[	[9.62, 15.21[	[15.21, 20.79[	[20.79, 26.37]
EFD-Jenks	[4.04; 5.94]	]5.94, 9.26]	]9.26, 19.42]	]19.42, 26.37]
AVST	[4.04; 5.53[	[5.53, 13.65[	[13.65, 21.78[	[21.78, 26.37]
KMEANS	[4.04; 6.37[	[6.37, 12.3[	[12.3, 22.95[	[22.95, 26.37]

TABLE III. POPULATION OF EACH BIN OF SAMPLE SX.

Method	Bin_1	Bin_2	Bin_3	Bin_4
EWD	6	0	3	3
EFD-Jenks	3	3	3	3
AVST	2	4	4	2
KMEANS	3	3	4	2

As it is easy to understand, we cannot find two discretization methods producing the same set of bins. As a consequence, the distribution of the values of SX is different depending on the method used.

B. Discretization Methods and Statistical Characteristics

When attempting to find the most appropriate discretization method for a column, what is important is not the law followed by its distribution, but the shape of its density function. This is why we first perform a descriptive analysis of the data in order to characterize, and finally to classify, each column according to normal, uniform, symmetric, antisymmetric or multimodal distributions. This is done in order to determine what discretization method(s) may apply. Let us underline that the proposed tests have to be performed in the given order:

1) We use the Kernel method presented in [15] to characterize multimodal distributions. The method is based on estimating the density function of the sample by building a continuous function, and then calculating the number of peaks using its second derivative. It involves building a continuous density function, which allows us to approximate automatically the shape of the distribution. The multimodal distributions are those which have a number of peaks greater than 1.
2) To characterize antisymmetric distributions in a next step, we use the Skewness, noted γ_3:

$$\gamma_3 = E[(\frac{X - \mu}{\sigma})^3] \quad (4)$$

The distribution is symmetric if $\gamma_3 = 0$. Practically, this rule is too exhaustive, so we relaxed it by imposing limits around 0 to set a fairly tolerant rule which allows us to decide whether a distribution is considered antisymmetric or not. The associated method is based on a statistical test. The null hypothesis is that the distribution is symmetric.
Consider the statistic: $T_{Skew} = \frac{N}{6}(\gamma_3^2)$. Under the null hypothesis, T_{Skew} follows a law of χ^2 with one

degree of freedom. In this case, the distribution is antisymmetric if $\alpha = 5\%$ if $T_{Skew} > 3.8415$.

3) We use then the normalized Kurtosis, noted γ_2, to measure the peakedness of the distribution or the grouping of probability densities around the average, compared with the normal distribution. When γ_2 is close to zero, the distribution has a normalized peakedness:

$$\gamma_2 = E[(\frac{X-\mu}{\sigma})^4] - 3 \quad (5)$$

A statistical test is used again to automatically decide whether the distribution has normalized peakedness or not. The null hypothesis is that the distribution has a normalized peakedness.

Consider the statistic: $T_{Kurto} = \frac{N}{6}(\frac{\gamma_2^2}{4})$. Under the null hypothesis, T_{Kurto} follows a law of χ^2 with one degree of freedom. The null hypothesis is rejected at level of significance $\alpha = 0.05$ if $T_{Kurto} > 6.6349$.

4) To characterize normal or uniform distributions, we use the Kolmogorov-Smirnov test, which can be used to compare the empirical functions of two samples if they have the same distribution.

These four successive tests allow us to characterize the shape of the (density function of the) distribution of every column. Combined with the main characteristics of the discretization methods presented in the last section, we get Table IV: This summarizes which discretization method(s) can be invoked depending on specific column statistics.

TABLE IV. APPLICABILITY OF DISCRETIZATION METHODS DEPENDING ON THE DISTRIBUTION'S SHAPE.

	Normal	Uniform	Symm-etric	Antisym-metric	Multimodal
EWD	*	*	*		
EFD-Jenks	*	*	*	*	*
AVST	*				
KMEANS	*	*	*	*	*

Example 2: Continuing Example 1, the Kernel Density Estimation method [15] is used to build the density function of sample SX (*cf.* Figure 2).

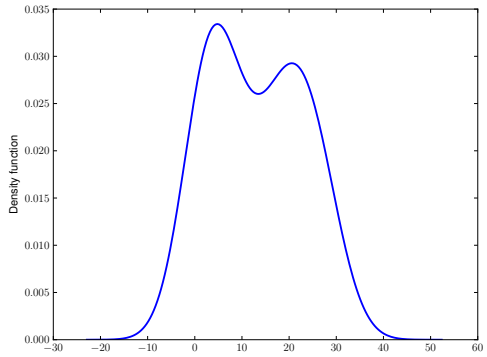

Figure 2. Density function of sample SX using Kernel Density Estimation.

As we can see, the density function has two modes, is almost symmetric and normal. Since the density function is multimodal, we should stop at this point. But as shown in Table IV, only EFD-Jenks and KMEANS produce interesting results according to our proposal. For the need of this example, let us perform the other tests. Since $\gamma_3 = -0.05$, the distribution is almost symmetric. As mentioned in (2), it depends on the threshold fixed if we consider that the distribution is symmetric or not. The distribution is not antisymmetric because $T_{Skew} = 0.005$. The distribution is not uniform since $\gamma_2 = -1.9$. As a consequence, $T_{Kurto} = 1.805$, and we have to reject the uniformity test. The Kolmogorov-Smirnov test results indicate that the probability that the distribution follows a normal law is 86.9% with $\alpha = 0.05$. Here again, accepting or rejecting the fact that we can consider if the distribution is normal or not depends on the fixed threshold.

C. Multi-criteria Approach for Finding the Most Appropriate Discretization Method

Discretization must keep the initial statistical characteristics so as the homogeneity of the intervals, and reduce the size of the final data produced. So, the discretization objectives are many and contradictory. For this reason, we chose a multi-criteria analysis to evaluate the available applicable methods of discretization. We use three criteria:

- The entropy H measures the uniformity of intervals. The higher the entropy, the more the discretization is adequate from the viewpoint of the number of elements in each interval:

$$H = -\sum_{i=1}^{NbBins} p_i \log_2(p_i) \quad (6)$$

where p_i is the number of points of interval i divided by the total number of points (N), and $NbBins$ is the number of intervals. The maximum of H is computed by discretizing the attribute into $NbBins$ intervals with the same number of elements. In this case, H reduces to $log_2(NbBins)$.

- The index of variance J, introduced in [10], measures the interclass variances proportionally to the total variance. The closer the index is to 1, the more homogeneous the discretization is:

$$J = 1 - \frac{\text{Intra-intervals variance}}{\text{Total variance}}$$

- Finally, the stability S corresponds to the maximum distance between the distribution functions before and after discretization. Let F_1 and F_2 be the attribute distribution functions before and after discretization respectively:

$$S = sup_x(\left|F_1(x) - F_2(x)\right|) \quad (7)$$

The objective is to find solutions that present a compromise between the various performance measures. The evaluation of these methods should be done automatically, so we are in the category of *a priori* approaches where the decision-maker intervenes just before the evaluation process step.

Aggregation methods are among the most widely used methods in multi-criteria analysis. The principle is to reduce

to a unique criterion problem. In this category, the weighted sum method involves building a unique criterion function by associating a weight to each criterion [16][17]. This method is limited by the choice of the weight, and requires comparable criteria. The method of inequality constraints is to maximize a single criterion by adding constraints to the values of the other criteria [18]. The disadvantage of this method is the choice of the thresholds of the added constraints.

In our case, the alternatives are the 4 methods of discretization, and we discretize automatically columns separately, so the implementation facility is important in our approach. Hence the interest in using the aggregation method by reducing it to a unique criterion problem, by choosing the method that minimizes the Euclidean distance from the target point $(H = log_2(NbBins),\ J = 1,\ S = 0)$.

Definition 1: Let D be an arbitrary discretization method, and V_D a measure of segmentation quality using the proposed multi-criteria analysis:

$$V_D = \sqrt{(H_D - log_2(NbBins))^2 + (J_D - 1)^2 + S_D^2} \quad (8)$$

The following proposition is the main result of this article: It indicates how we chose the most appropriate discretization method among all the available ones.

Proposition 1: Let DM be a set of discretization methods; the one, noted $\mathbb{D}$, that minimizes V_{D} (*see* equitation 8), $\forall D \in \{DM\}$, is the best discretization method.

Corollary 1: The most appropriate discretization method $\mathbb{D}$ can be obtained as follows:

$$\mathbb{D} = argmin(\{V_{\mathrm{D}}, \forall D \in \{DM\}\}) \quad (9)$$

As a result of corollary 1, we propose the MAD (Multi-criteria Analysis for finding the best Discretization method) algorithm, see Figure 3.

Input: X set of numeric values to discretize, DM set of discretization methods applicable
Output: best discretization method for X
1: **for each** method $D \in$ DM **do**
2: Compute V_D
3: **end for**
4: **return** $argmin(V)$

Figure 3. MAD: Multi-criteria Analysis for Discretization

Example 3: Continuing Example 1, Table V shows the evaluation results for all the discretization methods at disposal. Let us underline that for the need of our example, all the values are computed for every discretization method, and not only for the ones which should have been selected after the step proposed in Section IV-B (*cf.* Table IV). The results show that EFD-Jenks and KMEANS are the two methods that obtain the lowest values for V_D. The values got by the EWD and AVST methods are the worst: This is consistent with our optimization proposed in table IV, since the sample distribution is multimodal.

TABLE V. EVALUATION OF DISCRETIZATION METHODS.

	H	J	S	V_{DM}
EWD	1.5	0.972	0.25	0.313
EFD-Jenks	2	0.985	0.167	0.028
AVST	1.92	0.741	0.167	0.101
KMEANS	1.95	0.972	0.167	0.031

V. EXPERIMENTAL ANALYSIS

In this section, we present some experimental results by evaluating three samples. We decided to implement it using the MineCor KDD Software [4], but it could have been with another one (R Project, Tanagra, *etc.*) Sample_1 is a randomly generated file that contains heterogeneous values. Sample_2 and Sample_3 correspond to real data representing measurements provided by a microelectronics manufacturer (STMicroelectronics) after completion of the manufacturing process. Table VI sums up the characteristics of the samples.

TABLE VI. CHARACTERISTICS OF THE DATABASES USED.

Sample	Number of columns	Number of rows	Type
Sample_1	9	468	generated
Sample_2	7	727	real
Sample_3	1281	296	real

Figures 4 and 5 summarize respectively the evaluation of the methods used on the two first samples.

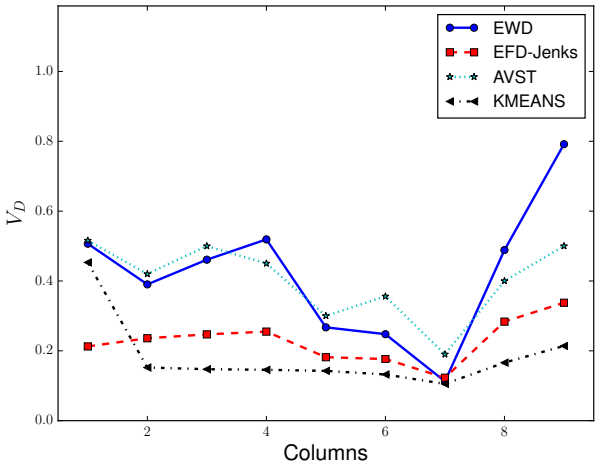

Figure 4. DM comparison on Sample_1's columns.

For the Sample_1 evaluation shown graphically in Figure 4, the columns studied have relatively dispersed, asymmetric and multimodal distributions. "Best" discretizations are provided by EFD-Jenks and KMEANS methods. We note also that the EWD method is fast, and sometimes demonstrates good performances in comparison with the EFD-Jenks or KMEANS methods.

For Sample_2 attributes, which have symmetric and normal distributions, the evaluation on Figure 5 shows that the EFD-Jenks method provides generally the best results. The

 ISBN: 978-1-61208-415-2

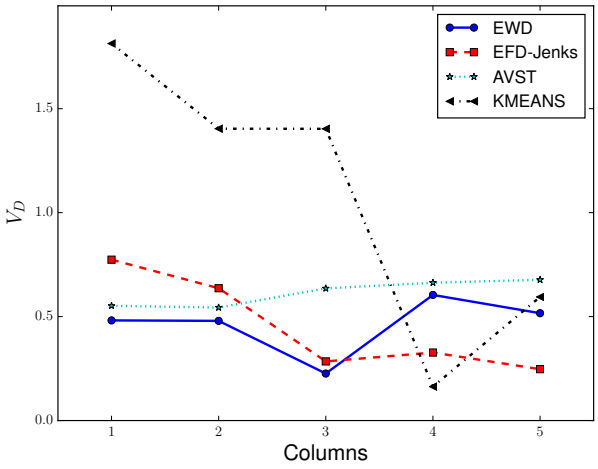

Figure 5. DM comparison on $Sample_2$'s columns.

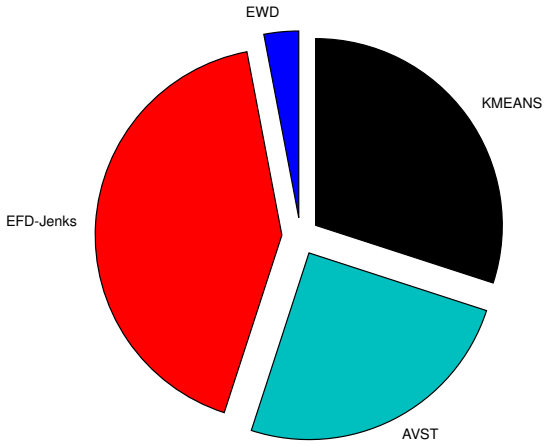

Figure 6. Selected Discretization Method.

KMEANS method is unstable for these types of distributions, but sometimes provides the best discretization.

Finally, Figure 6 summarizes our approach: We have tested it over each column of each dataset. Any of the available methods is selected at least once in the dataset of the three proposed samples, which enforces our approach. As expected, EFD-Jenks is the method that is the most often kept by our software ($\simeq$ 42%). AVST and KMEANS are selected approximately 30% each. EWD is only selected a very few times (less than 2%).

VI. Conclusion and Future Work

In this paper, we presented a new approach for automatic data preparation implementable in most of KDD systems. This step is generally split into two sub-steps: (i) detecting and eliminating the outliers, and (ii) applying a discretization method in order to transform any column into a set of clusters. In this article, we show that outliers' detection is depending on if data distribution is normal or not. As a consequence, we do not have to apply the same pruning method (Box plot *vs.* Grubb's test). Moreover, when trying to find the most appropriate discretization method, what is important is not the law followed by the column, but the shape of its density function. That is why we propose an automatic choice for finding the best discretization method based on a multi-criteria approach. Experimental evaluations done on real and synthetic data validate our work, showing that it is not always the very same discretization method that is the best: Each method has its strengths and drawbacks.

For future works, we aim to experimentally validate the relationship between the distribution shape and the applicability of used methods, to add other discretization methods (Khiops, Chimerge, Entropy Minimization Discretization, *etc.*) to our system, to parallelize our work using the latest functionalities of multicore programming, and to measure the impact of the data preparation step on the results of some mining algorithms (association rules, correlation rules, *etc.*).

References

[1] D. Pyle, Data Preparation for Data Mining. Morgan Kaufmann, 1999.

[2] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 7, 2002, pp. 881–892.

[3] C. Aggarwal and P. Yu, "Outlier detection for high dimensional data," in SIGMOD Conference, S. Mehrotra and T. K. Sellis, Eds. ACM, 2001, pp. 37–46.

[4] C. Ernst and A. Casali, "Data preparation in the minecor kdd framework," in IMMM 2011, The First International Conference on Advances in Information Mining and Management, 2011, pp. 16–22.

[5] O. Stepankova, P. Aubrecht, Z. Kouba, and P. Miksovsky, "Preprocessing for data mining and decision support," in Data Mining and Decision Support: Integration and Collaboration, K. A. Publishers, Ed., 2003, pp. 107–117.

[6] M. Grun-Rehomme, O. Vasechko et al., "Méthodes de détection des unités atypiques: Cas des enquêtes structurelles ukrainiennes," in 42èmes Journées de Statistique, 2010.

[7] J. W. Tukey, "Exploratory data analysis. 1977," Massachusetts: Addison-Wesley, 1976.

[8] F. E. Grubbs, "Procedures for detecting outlying observations in samples," Technometrics, vol. 11, no. 1, 1969, pp. 1–21.

[9] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," Journal of the American Statistical Association, vol. 62, no. 318, 1967, pp. 399–402.

[10] C. Cauvin, F. Escobar, and A. Serradj, Cartographie thématique. 3. Méthodes quantitatives et transformations attributaires. Lavoisier, 2008.

[11] I. Kononenko and S. J. Hong, "Attribute selection for modelling," Future Generation Computer Systems, vol. 13, no. 2, 1997, pp. 181–195.

[12] S. Kotsiantis and D. Kanellopoulos, "Discretization techniques: A recent survey," GESTS International Transactions on Computer Science and Engineering, vol. 32, no. 1, 2006, pp. 47–58.

[13] U. of Kansas. Dept. of Geography and G. Jenks, Optimal data classification for choropleth maps, 1977.

[14] A. K. Jain, "Data clustering: 50 years beyond k-means," Pattern Recognition Letters, vol. 31, no. 8, 2010, pp. 651–666.

[15] B. W. Silverman, Density estimation for statistics and data analysis. CRC press, 1986, vol. 26.

[16] P. M. Pardalos, Y. Siskos, and C. Zopounidis, Advances in multicriteria analysis. Springer, 1995.

[17] B. Roy and P. Vincke, "Multicriteria analysis: survey and new directions," European Journal of Operational Research, vol. 8, no. 3, 1981, pp. 207–218.

[18] M. Chilali, "Méthodes lmi pour l'analyse et la synthèse multi-critère." Ph.D. dissertation, 1996.

 ISBN: 978-1-61208-415-2