CHAPTER 4

STANDARDIZATION ALGORITHM

4.1 GENERAL

Data preprocessing techniques are applied to a raw data to make the data clean, noise free and consistent [Vaishali and Rupa, 2011]. Ranked lists are encountered in research and daily life, and it is often of interest to compare given set of lists, even when they are incomplete or have only some members in common to identify the most suitable thing to a particular application. Interesting set of association rules will selected in the post processing of KDD by fixing the threshold on interestingness score (IS score) on a data set by IMs. Hence selection of suitable IM plays a vital role in the post processing of KDD, Towards the selection of alist of appropriate measures by score “Standardization Algorithm” was introduced in this chapter.

4.2 WHAT IS STANDARDIZATION?

In general raw data having more useful and interesting knowledge, but the outliers and noises are dominating and occupying the valuable knowledge places. Hence making the data as a quality one by removing those values leads to better results. A simple way of removing outliers may do by converting the data into specific range. This can be done applying standardization techniques.

In data mining standardization plays as a central preprocessing step. Also widely used by researchers to handle contingency tables of varying marginal value.Tan *et al.* [2004] selected the right objective measure for association by standardization and confirmed that this technique is useful in getting a better idea of knowing underlying association between the variables. They also proved that IMs becomes consistent by standardizing contingency table of association rule. So applying standardization on score may lead to consistent of measures.

4.2.1 Methods of Standardization

Standardization can be done in many ways, for numerical databases with attributes having Gaussian distributions, the most common procedure applied to transform the attribute x with zero mean and unit variance by Z –score standardization defined below

$$Z = \frac{x - \mu}{\sigma} \tag{4.1}$$

μ- mean, σ- standard deviation.

This can be done by min-max and decimal scaling methods. Mohamad and Usman[2013] compared the above said methods on k-means clustering algorithm and proved that, Z -score is having better performance. It has been proved by many researchers[Tan *et al.,* 2004. Alberchen *et al.*, 2006, Li *et al.*, 2011 and Panichkitkosolkul, 2013] standardization make the data perfect and includes all the dates into algorithms will comparatively give better performance.

4.2.2 Limitations

Applying standardization of data will bring the data closer by removing outliers, hence converts their range to a specific interval. Data becomes a dimensionless by applying this technique. This process used to define standard indices that make the original data into standard data by losing its location of knowledge and scale. Hence, applying standardization will convert the scores of different measures defined by researchers in different situations to a specific range.

4.3 COEFFICIENT OF VARIATION

While analyzing a set of data or performance measures on a dataset, comparison of numerical distributions derived on different scale becomes leased step. To do the comparison in a better way statistically a dimensionless measure called Coefficient of Variation (CV) has been used widely.

[Alberchen *et al.*, 2006 and Tian, 2005].They are also stating that, CVis used as a measure of precision for calculating the variation of data set. Ostle[1988] and Panichkitkosolkul [2013]confirm that the variation of a measure measured by two different formulae remains same. For example the variation of temperature measured on two different unit systems Celsius and Fahrenheit will remains same.And the same CV may compare the dispersion of two or more members of series measured in different units and also that series with the same units, but running at different level of magnitude. Similarly, the CVs have been used to evaluate results from different experiments involving the same units of measure, possibly conducted by different persons.Hence comparing the scores obtained by different measures on a dataset can be madeby CV, which is defined as follows,

The Coefficient of Variation (CV) of a variable X of having adistribution function with mean μ and variance σ^2 is defined by

$$Coefficient\ of\ Variation = \frac{\sigma}{\mu} X100 \qquad (4.2)$$

Li *et al.*, [2011] confirms that CV functioning as a factor on selecting appropriate interestingness measure. CV is a special case of analyzing variability applied in this research with the meaning of variability.

Statistically, it is the fact that lower the CV leads, the less deviation among the variables and higher the CV leads to more deviation among the variables. Hence lower the CV leads to less variation among scores, hence the measures having less CV value tends to produce more interesting rules. CV has been applied as a better predictor of variability among datasets by many researchers, also is often used to compare the variability of two or more groups [George *et al.*,2002, Mahumouduan et al., 2007, Xiong and Chen, 2006, Mohsen, 2010, Weber, 2010, Li *et al.*, 2011and Golay *et al.*, 2013].

4.3.1 Coefficient of Variation – Why?

The existing variability measures are dependent of mean value, hence variability measured by those measures may change whenever the mean value changes. That is range of those measures score becomes larger when the mean is larger, and become smaller when the mean is smaller. Hence, selecting appropriate IM by comparing their scores may have the following issues (i) A measure of high (low) range will be selected always. That is a measure having either high or low score will dominants the selection. (ii) Measures score of other end will not be selected.

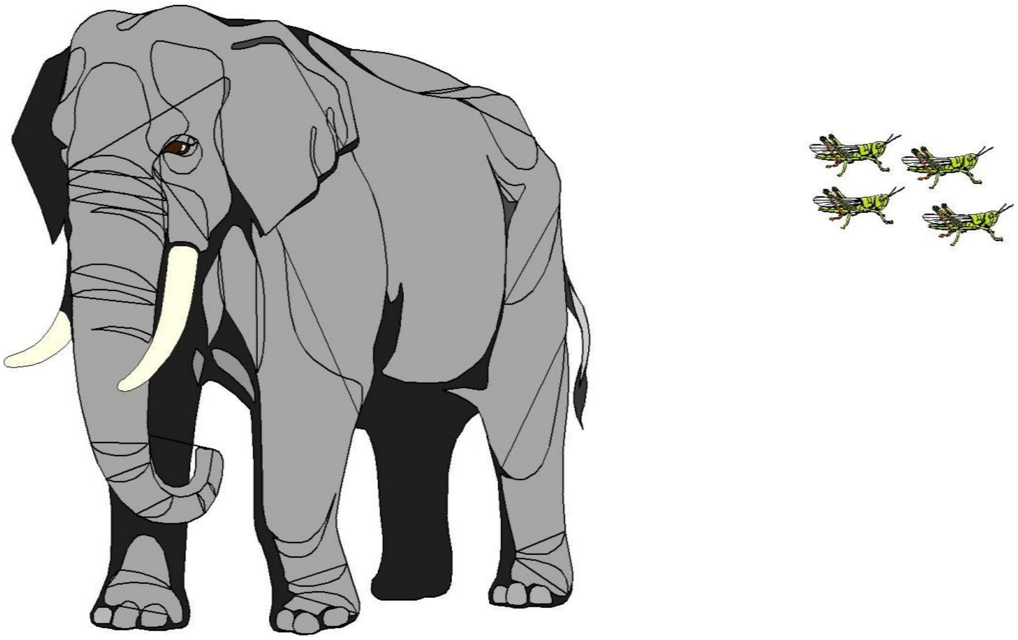

Fig. 4.1. Before Standardization

To avoid the above said issues comparison may done by converting the scores into specific range. In this work CV is applied to convert the scores into specific range.

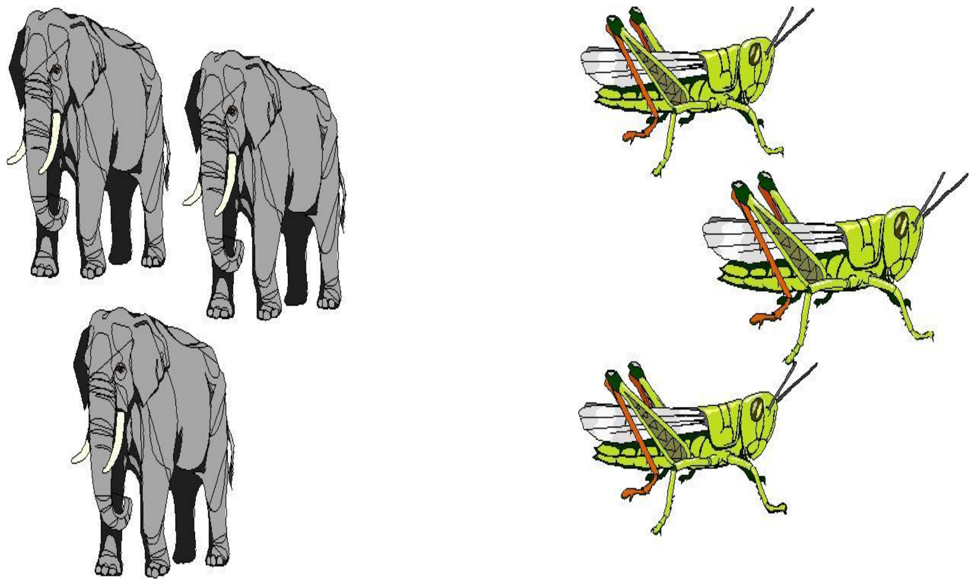

Fig. 4.2.After Standardization

To know the interesting fact about CV, let us have the following discussion. Is the size of Elephants are “more variable” than the size of grasshoppers? Represented in Fig. 4.1. The answer is absolutely yes! But this decision needs to be analyzed with the effect of the mean.

Let’s put the elephants on a diet (or fatten up the grasshoppers) until each group has the same mean size then see Fig.4.2. Now are the sizes of Elephants ‘more variable’ than the sizes of grasshoppers when the means are ‘made’ equal? The answer is not obvious! Instead of fattening up the grasshoppers or putting the elephants on a diet, the same result may obtained by dividing every value in the populations by its mean.

Variability of an analytic procedure producing continuous values was not summarized by the SD but done by CV, because CV is a ratio of SD to the average, and this result can obtain as a percentage [George *et al.*,2002]. Also, it is the fact that the SD value of a set of scores increase (decrease) as the average value of the score increases (decreases), so dividing the SD by average will removes averageas a factor in calculation of variability, this is the fact in applying CV instead of SD, while comparing two series.The process discussed above is a standardized process, also standardizingof SD leave space to compare the variability of estimation. That is, a use of CV for the comparison of variability among series of data will lead meaning full outcomes.

4.3.2 Limitations

Applying CV has two advantages; it is dimensionless, therefore do not vary as its measurement unit changes. Secondly,CV may define the results if the variability of a current process exceeds by the values obtained from past performances, so CV plays the role of quality control. Even though CV has these advantages, it has the following disadvantages.

It predicts wrong deviation, when the variables having negative values or the average of the variable close to zero. And we know that if we measure temperature by Celsius and Fahrenheit units, the variation between Celsius and Fahrenheit units remains the same. Martinez Pons[2013] stated that, coefficient of variation used to compare two standard deviations, when their averages differ substantially, and its value becomes larger when variance becomes greater than the mean and in this case size of CV is impossible. Due to the advantages as well as the disadvantages of CV, it should be used with care [Weber, 2010].

4.4 STANDARDIZATION ALGORITHM

Different IMs in the literature is having scored in different range. Hence, to rank the IMs transforming their scores into specific range is necessary. By the facts studied above, average free standardized score will be obtained by applying CV. Therefore, by applying CV a ‘Standardization Algorithm’ was proposed to rank IMs.

In general, IMs existing in literature is equivalent with one or more number of measures, derived from other measures (Most of them from basic measures) or statistically defined. It is clear studied by the previous chapters different measures having the different score range, this will not lead to unique ordering while comparing by range. Also ARM has the advantage of allowing an unsupervised extraction of rules and of illustrating implicative tendency in data: It has the advantage of producing a prohibitive number of rules. In rule evaluation, main difficulties faced are how to extract an interesting set of rules from the huge amount of extracted rules. And the proposal of many interestingness measures in the literature leads to another difficulty that, how to select IMs, that are adapted to its goal and its data, to extract the most interesting set of rules.

And the basic measures in the pattern evaluation are support, confidence and lift. But each one of these has some drawbacks. Heravi and Zaiane[2010] stated that in case of choosing large minimum support leads only to the rules, that contain obvious knowledge and missing the expectation case that are interesting. Whereas, choosing a low minimum support produces so many rules which could be redundant and noisy. Confidence is also not a perfect measure since it produces confident association between the statistically independent items. Similarly, lift leads to wrong perdition in correlation that is in case of negative correlation it shows positive correlation, because the lift is not depending on the null records.

Towards the elimination of the above said difficulties, standardization Algorithm was established based on CV by considering top most rules of extracted association rules. It is proved that generally small CV will provide high classification accuracy, and directly related to its range [Li *et al.,* 2011]. In general, rules with large IS scores are considered as most important compare to other rules with small scores, even though they are evaluated over the significance threshold level.

Tan *et al.* [2004] listed that, before deciding right measure to a particular domain, the user must analyze several key factors, in this continuation, our algorithm decides perfect measures for a dataset based on the variation projected by CV value. Geng and Hamilton[2006] suggested a promising method to find the interestingness using the automatic selection or combining appropriate measures. Khan and Sheel[2013] also stated the importance of auto selection of their computing system for analysis of DNA sequences using OPTSDNA algorithm. Our algorithm ensures the automatic selection of measures. Hiep [2010] stated number of interestingness measures which may be reduced by considering a common measure on two or more measures. Also the interestingness of a measure can be calculated by the participating measures on a measure.

4.4.1 Measurement Matrix or Score Matrix

Let us consider, M number of Top most association rules on the data set $\mathcal{C}$, which are mined by data mining algorithm, Represent each measure on the Row and K number of measures are appropriately suitable to evaluate the rules. Each measure is represented asa column. Then k measure's IS score on M rules are represented by M×K matrix is known as score matrix. This has been represented in appendix I, in the datasets 1 to 5, Cleveland, Breast, Leaf and Dress sale. These scores may helpful in defining the interestingness of measures with respect to the expected knowledge present on the given dataset.

4.4.2 Not Applicable Measure

A measure M_k is said to be Not Applicable (NA) measure to a data set if the average score fall into the range close to zero or the score standard deviation of M_k is large compared to the average score. A set of measures not satisfying the above definition are called applicable measures. These sets of measures mayinclude for further analyze to identify the good measures.

4.4.3 CV of a dataset

Let $\mathcal{C}$ be a data set and M_k is the measure, then the CV of thedataset $\mathcal{C}$ with respect to the measure M_k is defined by

$$CV_k(\mathcal{C}) = \frac{\sigma_k(\mathcal{C})}{A_k(\mathcal{C})} X100 \tag{4.3}$$

$A_k(\mathcal{C})$, the average IS score of $\mathcal{C}$ with respect to the measure M_k, defined in section 5.1

4.4.4 Equivalent Measures

Two measures M_k and M_l are said to be equivalent if

$$CV_k(\mathcal{C}) = CV_l(\mathcal{C})\ for\ k \neq l \tag{4.4}$$

Two or more numbers of measures with same CV value are called equivalent set of measures. Identifying these kinds of measures will helpful in the reduction of measures.

4.4.5 Ordering Principle

The given set of measures will be classified and ordered as follows by the factor $A_k(\mathcal{C})\ and\ CV_k(\mathcal{C})$

- A set of measures having $A_k(\mathcal{C})$ in the range close to zero or $\sigma_k(\mathcal{C}) > A_k(\mathcal{C})$ are grouped as not applicable measures
- The remaining set of measures are called applicable set of measures
- Two or more numbers of measures having the same CV value are grouped as equivalent set of measures
- A measure M_k is earlier than M_l if

$$CV_k(\mathcal{C}) < CV_l(\mathcal{C}) for\ k \neq l \tag{4.5}$$

4.4.6 Algorithm Description

The ordering principle described above may helpful in deciding measures if it formulated as an algorithm. The top most set of association rules of the form 2×2 Contingency tables $C_{1,}$ $C_{2,}$ $C_{3,}$ … $C_{i,}$ $C_{i+1,}$ … C_m and set of measures are given as input. Initially, each $C_{i,} \epsilon\ C$ transformed into IS score equivalent to the measures M_k and listed as k column vectors. The collection of k column vectors represented as measurement matrix M and the order of matrix is given by $m \times k$(number of association rules × number of measures). Each column in the measurement matrix is the numerical equivalent of top most association rules with respect to the data set. The average of each column k is calculated by the equation. The measures of columns whose average IS score zero and close to it are listed as set of not applicable measures. The rest of the measures are considered as applicable measures. For columns having applicable measures, standard deviation (σ) will be calculated. By applying mean and standard deviation value in the Equation (4.3)will yield Coefficient of Variation (CV) value to the respective measures. These measures are arranged by the increasing order of CV. Thus we obtained the ranking of measures from most suitable to least. That is the measure having less CV leads to a perfect measure.

Variations obtained by applying the standardization algorithm will reduces the score range considerably. That is all the scores into a unique range since the CV value obtained by dividing the deviation from average by SD.

4.4.7 Steps in identifying Perfect Measures

The perfect measure of interestingness identified by applying the following steps by the standardization algorithm.

- Consider the top set of top association rules in the form of contingency table mined from a dataset D.
- Consider a set of measures to be ranked possibly having n number of measures.
- Obtain a measurement matrix, by considering rules on row and measures on columns.
- For each measure $m_k \in M$ and for each AR $c_i \in C$ calculate the measure score. That is the score of m_k on c_i and it will be represented as $m_k(c_i)$.
- Represent the measures scores of n measures on m rules as a mxn matrix. Now the set of scores of k^th measure will be represented by column k. for the set of n measures the measures scores are represented by n columns.

Standardization Algorithm

```
Input: Association rules of the form 2×2 contingency table and set of measures
$M_1$, $M_2$, M3…, $M_{k,}$ $M_{k+1,…}$$M_n$
Output:
    • Ascending order of Applicable measures.
    • Set of Measures not applicable.
Algorithm:
    1. Get set of 2×2 contingency tables $C_1$, $C_2$, $C_{3,}$ …, $C_{i,}$ $C_{i+1,…}$Cm
    2. Get set of measures M = { $M_1$, $M_2$, M3…, $M_{k,}$ $M_{k+1,…}$$M_n$ }
    3. For i = 1 to m and for k = 1 to n Compute $M_k(C_i)$
    4. Represent $M_k(C_i)$ as a Matrix M ={$M_{ik}$}, where i=1 to m and k = 1 to n.
    5. Find Mean of each column k, A(k),
    6. List K values for which A(K) = 0
    7. Remove the columns having A(k) = 0
    8. List the Measures having A(k) = 0
    9. Calculate Coefficient of variation $CV_k$ for each column k, for k= 1 to n
    10. Sort by ascending order of $CV_k$
    11. End
```

Fig.4.3. Standardization Algorithm

- For each column k the mean score $A(k)$ and standard deviation $\sigma_k(\mathcal{C})$ are calculated.
- The columns with $A(k) \in (-0.1,\ 0.1)$ or $\sigma_k(\mathcal{C}) > A_k(\mathcal{C})$ are eliminated, and the measures of respective columns are listed as not applicable measures.
- For the remaining columns $CV_k(\mathcal{C})$ values are obtained

- If $CV_k(\mathcal{C})$ is same for two or more number of columns, those column measures are listed as equivalent measures.
- Sort the rest of the column measures by the ascending order of $CV_k(\mathcal{C})$. In the sorted list the top most measures will be considered as a perfect measure.

Finally, A list of measures are obtained by the CV value. Variance is a special kind of variability. Hence, the variance finded by applying the standardization algorithm will be equivalent as a variability value here. By considering the above steps a standardization algorithm of ranking were presented in Fig. 4.3. A measure of least ranking obtained from the set of measures by this algorithm was called 'perfect measure' with respect to the dataset D.

4.5 SUMMARY

This chapter proposes a ranking algorithm called standardization algorithm, which functioning by converting variations of scores into a specific range [0, 100]. Equivalent measures, applicable and not applicable measures are identified finally applicable measures are ranked according to the variation measured, least ranked measure identified as a perfect measure. Even though producing a better list of measures the elimination of some measures as not applicable, considered as a drawback and leads to the further analysis.