Journal of Engineering and Applied Sciences 9 (7): 249-253, 2014
ISSN: 1816-949X

Metrics Free Techniques and Issues to Acquire Unifeatured High Density Quality Clusters

[1]G. Abel Thangaraja, [2]Saravanan Venkataraman Tirumalai and [3]A. Pankaj Moses Monickaraj
[1]Department of Computer Science, Kaypeeyes College of Arts and Science, Kotagiri, India
[2]College of Computer and Information Sciences, Majmaah University, Majmaah, Kingdom of Saudi Arabia
[3]Department of Computer Science, Bharathiar Univeristy, Coimbatore, India

Abstract: There are various metrics to measure the efficiency of performance say for memory byte, kilobyte and megabyte, for time, milli and micro second. Of the various research domains in data mining, clustering the unsupervised classification is one of unique area for research. To call a cluster with better quality, the intra clustering similarity should be minimum and inter clustering density, similarity should be maximum. In this study, few of the issues and techniques that have to be focused on to acquire unifeatured high density quality clusters are elaborated along with a statistical approach. The entire research study primarily focuses by 8 dimensions which are categorized into 4 each for techniques and methods.

Key words: Data mining, cluster quality, metrics, techniques and methods, inter clustering density

INTRODUCTION

Data mining: Data mining is also called as Knowledge Discovery in Databases (KDD) which can be defined, as the extraction of implicit, previously unknown and potentially useful knowledge patterns from huge amount of data. Data mining specifically refers to extracting knowledge from large amounts of data through its various branching like classification, clustering, etc. Data mining techniques are used in different application to analysis and predict the data for decision support system (Jeyabalaraja and Prabakaran, 2012).

Clustering: Cluster is a collection of data objects which are similar to one another within the same cluster and dissimilar to the objects in other cluster. Finding similarities between data, according to the characteristics found in the data and grouping similar data objects into clusters is called as cluster analysis. Ensemble analysis improves classification accuracy and the general quality of cluster solution (Strehl and Ghosh, 2002). Clustering is a vital process to distinguish the objects, according to either their shapes or characters. Clustering is used as unsupervised learning process (Rai and Singh, 2010) and its goal is to discover a new set of categories which is one of the most useful tasks in the data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data (Jeyabalaraja and Prabakaran, 2012). Thus, the main concern in the clustering process is to cluster sensible groups which allow us to discover similarities and differences, as well as to derive useful inferences about them. Clustering is also called as unsupervised classification.

Role of metrics in data mining: Data mining has been the potential technology for supporting, enhancing and understanding data. The absence of the metric leads to redundancy while producing clusters. Irrespective of the research domain the efficiency of each process performance can be measured by means of metrics such that productive relationship among features can be modeled. Intelligent computing techniques, such as data mining can be applied in the study of software quality by analyzing software metrics (Yang *et al*., 2006). Not much research has carried out in using the metrics in data mining. If technological metrics are used extensively, the knowledge discovered from the data mining tools would be of better quality. The proposed research work considers this technologies and methods that have to be taken into account for effective clustering.

Role of metrics in cluster quality: The best method in predicting software quality is dependent on practical dataset and clustering analysis technique has advantages in software quality prediction, since it can be used in the

Corresponding Author: G. Abel Thangaraja, Department of Computer Science, Kaypeeyes College of Arts and Science, Kotagiri, India

case having little prior knowledge (Yang *et al*., 2006). The role of software clustering is to identify concrete entities for which a mapping decision is easy enough to be made automatically. The goal of clustering was to identify related components in the software system (Shtern and Tzerpos, 2012). Effective usage of the following specification on clustering techniques leads to quality clusters.

TECHNIQUES AND ISSUES

Irrespective of the algorithm used the following technological and issues have to be more stressed on to obtain and fine tune the results effective during clustering (Jeyabalaraja and Prabakaran, 2012).

Technology is an application of science (the combination of the scientific method and material) to meet an objective or solve a problem. For the clustering, the most dominant 4 techniques are focused in this study.

- Cluster-size, validity
- Complexity
- Coupling
- Cohesion

In parallel to few issues:

- Cost (time involved in cluster formation, repositioning, hardware, familiarity of the mining expert and overheads)
- Factors (quality of mining expert, experience and staff attrition ratio)
- Quality (by considering noise and rate at which faults are found)
- Maintainability

are also focused on in this study.

Cluster-size, validity: The size may either increase or decrease according to the size of the data and the working pattern of the algorithm. The growth of the cluster should grow gradually. In data mining, as the database contains historical data, the size of the cluster is considered to be an issue for the researchers. A well defined mathematical model/framework is the need of the hour.

The size of each cluster after fixing the number of clusters can be obtained by using either Gompertz function or binomial distribution (Gupta and Kapoor, 2010).

Gompertz function:

$$y(t)=ae^{be^{ct}}$$

This method is applicable only when the time factor is known and it is cumbersome to estimate the size of the clusters. For computation purpose, researchers use logarithm:

$$\log y(t)=\log a+be^{ct}\log e$$

Binomial distribution: The probability mass function of binomial distribution is:

$$p(x)=n_{c_x}p^xq^{n-x},x=0,1,2,...,n$$

Where:

p = Probability of selecting a cluster
q = Probability of not selecting a cluster
n = Number of clusters
N = The number of members in the graph
p (x) = The probability of selecting xth cluster
x = 0, 1, 2, ..., n
N*p (x) = The size of the xth cluster

To find the size of each cluster (Table 1). Consider p = 0.6, q = 0.4 (p+q = 1), n = 10 and N = 2000. It is noted that if one may assume the number of clusters n and total number of members N in different levels, he/she may get other sizes of clusters.

On the other end, evaluating and assessing the results of a clustering algorithm is known as cluster validity. It is mainly classified into 3 types (Halkidi *et al*., 2002):

- External validity criteria
- Internal validity criteria
- Relative validity criteria

External validity criteria measure how much the clustering results match the prior knowledge about the data. Internal validity criteria measure how well the clustering results match the future knowledge about the data. The earlier 2 validity criteria depend on statistical testing which needs high computational demands. Cluster validity helps to identify whether the particular cluster is of good quality.

Table 1: Size of clusters

X	n_{c_x}	p^xq^{n-x}	Size
0	1	$1.048*10^{-4}$	0
1	10	$1.573*10^{-4}$	3
2	45	$2.359*10^{-4}$	21
3	120	$3.539*10^{-4}$	85
4	210	$5.308*10^{-4}$	223
5	252	$7.963*10^{-4}$	401
6	210	$1.194*10^{-3}$	502
7	120	$1.792*10^{-3}$	430
8	45	$2.687*10^{-3}$	242
9	10	$4.031*10^{-3}$	81
10	1	$6.047*10^{-3}$	12

Complexity: Computational complexity problem is understood to be a task solved by a computer which is equivalent to stating that the problem may be solved by mathematical techniques. Computational complexity in clusters can be solved by means of landmark based dimensionality reduction algorithm which results in less cost and time. To achieve cluster quality, reduce the strength of computational complexity in clusters. In the absence of computational complexity metric, clustering the data will be too complex.

Algorithmic complexity is concerned about how fast or slow particular algorithm performs. Big Omega (or) Big O is one of the notation's to express an algorithm runtime complexity. The time function T(n) can be related with Big O. For example, the following statement:

$$T(n)=O(n^2)$$

Says that an algorithm has a quadratic time complexity. Further the Big Omega can be extended to:

- Constant time O(1)
- Linear time O(n)
- Logarithmic time O(log n)
- Quadratic time $O(n^2)$

Coupling: A coupled cluster consists of elements that are similar to one another and distinct from members in other clusters (Marx *et al*., 2002). Coupling measures the number of collaborations that a cluster has with any other clusters. Higher coupling decreases the reusability of a quality cluster. Higher coupling complicates modifications and testing. Coupling should be kept, as low as possible. If this coupling metric is not done, it is not possible to categorize the members which lead to absence f quality.

Coupling can metrically measured as low (also loose and weak) or high (also tight and strong) and few of the types of coupling are listed as:

- Pathological coupling
- Global coupling
- External coupling
- Control coupling
- Data-structured coupling
- Data coupling
- Message coupling
- Subclass coupling
- Temporal coupling

$$\text{Coupling}(C)=1-\frac{1}{d_{in}+2xc_{in}+d_{op}+2xc_{op}+g_d+2xg_c+w+r}$$

For data and control flow coupling:

d_{in} = Total number of input data parameters
c_{in} = Total number of input control parameters
d_{op} = Total number of output data parameters
c_{op} = Total number of output control parameters

For global coupling:

g_d = Total number of global variables used as data
g_c = Total number of global variables used as control

For environmental coupling:

w = Total number of modules
r = Total number of modules calling the module under consideration

Coupling (C) makes the value larger the more coupled the modul. If the number ranges from 0.67 (low coupling) to 1.0 (highly coupled).

Cohesion: Cohesion is a qualitative measure meaning that the source code text to be measured is examined using a rubric to determine a cohesion classification. Cohesion is may also be defined, as the sum of the weights of all links within a cluster. The cohesion of a class is the degree to which its set of properties is part of the problem or design domain. The best cohesive cluster is formed when all elements of a cluster belong to the same category (Sileshi and Gamback, 2009). The qualities of the clusters created by the algorithms are measured in terms of cluster cohesion. The types of cohesion, in order of the worst to the best type are as follows:

- Coincidental cohesion (worst case)
- Logical cohesion
- Temporal cohesion
- Procedural cohesion
- Communicational cohesion
- Sequential cohesion
- Functional cohesion (best case)

In the absence of cohesion metric, there is no chance for grouping the same category of members which finally leads to less quality.

Cost: The expenditure involved in collection and cleaning data, cluster formation, repositioning, hardware, familiarity of the mining expert, overheads and etc., is generally refereed as cost. Without cost one cannot imagine to produce quality cluster. The selection of quality cluster

with less cost is essential. The time required for the earlier stated factors are taken into account. Thus, the Total Cost (TC):

$$TC = C_c t_1 + C_f t_2 + (C_h + C_w) t_3 + C_m$$

The cost and time are symbolically stated as follows:

C_c = Cost for collection and cleaning data
C_f = Cost for formulation of clusters
C_h = Cost for hardware
C_w = Cost for mining expert
C_m = Miscellaneous costs
t_1 = Time duration of collection and cleaning the data
t_2 = Time duration of formulation of clusters
t_3 = Time duration of machine processing (hardware)

With this TC, the cost required for the performance of the algorithm can be obtained.

Factors: There are various factors while processing with issues, one of the most dominant is the Human Factor (HF) which is not that much considered but in this study, it is considered as the vital issue and has been focused on.

Quality of mining expert should have the attitude towards data mining, knowledge about the clusters, awareness of datasets, interest on studying data mining problems and knowledge to interpret results, etc., experience of mining expert will provide accurate and concrete results in lesser time to utilize future studies. Attrition ratio describes the rate at which employees leave a company. Staff attrition ratio must be very less and it will give quality clusters. If anyone of the human factors fails, the resulting cluster will not be of good quality. HF is a function of the sum of experience, psychological measures and attrition ratio.

$$HF = f(a, b, c) = f\left[a + \varphi(\alpha, \beta, \gamma, \delta) + c\right]$$

Let:
a = Number of years of experience
b = φ (α, β, γ, δ)-psychological measures
α = Attribute
β = Knowledge
γ = Awareness
δ = Interest
c = Attrition ratio

Quality: Quality means the standard of something measured against other things of a similar kind or the degree of excellence of something. Number of lines of code, load time, execution time, size of program (binary), modularity and density of the clusters are the key focus.

To improve the earlier said time, etc., the efficiency of the programming skill has to be improved to obtain better results.

Irrespective of the language used for clustering cyclomatic complexity, halstead complexities are the statistics metrics which focus on the quality of codes.

Cyclomatic approach: The cyclomatic approach counts the number of linearly independent paths through the source code or level of confidence in the program.

$$\text{Cyclomatic complexity } M = e - n + 2p$$

Where:
e = Number of edges of the graph
n = Number of nodes of the graph
p = Number of connected components

According to the number of decision making syntaxes, measurements can be done.

Halstead approach: Halstead reflects the implementation or expression of algorithms in different languages but it is always independent of their execution on a specific platform. The metrics are computed statically from the code so that their goal is to identify measurable properties and the relations among them. The difficulty measure is related to the difficulty of the program to write or understand. Let:

$$\text{Program vocabulary: } O = O_1 + O_2$$

$$\text{Program length} = N_1 + N_2$$

$$\text{Program length} = O_1 \log_2 O_1 + O_2 \log_2 O_2$$

$$\text{Volume} = N \log_2 O$$

$$\text{Difficulity} = \frac{O_1}{2} \times \frac{N_2}{O_2}$$

$$\text{Effort} = \text{Difficulity} \times \text{Volume}$$

Where:
O_1 = Number of distinct operators
O_2 = Number of distinct operands
N_1 = Total number of operators
N_2 = Total number of operands

Thus, Halstead complexity would provide the tentative results for the particular source code can be calculated by the earlier.

Maintainability: Maintainability metric can be used to maximize the cluster's long life, cluster efficiency, cluster reliability and minimize noise, detect faults. The produced cluster need not be perfect and it may fail during its operation due to time, cost. The ease with which repair and enhancement may be made to the cluster. If the produced clusters are not maintained properly, it leads to quandary. The proposed approach will take this issue into consideration during the prototype development.

The mathematical concepts like measures of central tendencies, dispersion measures, Laplace transforms, probability distributions, failure rates and reliability measures are used to achieve maintainability. The mathematical formulation of maintainability is not fixed model but it changes according to the concern flexibility of the research work which is used.

DISCUSSION

To sum up to and to lean on to a conclusion, let us pick the US army commission case study dimension of approach.

US army commissioned a study on how to redesign the uniforms for female soldiers. The army's goal was to reduce the number of different uniform sizes that have to be kept in inventory. Researchers designed a new sizes based on the actual shapes of women in the army. Instead of having several sizes for each soldier, they clustered on the body shapes (height, weight, short legged, small waist, large busted).

In Fig. 1, x-axis represent the body shape and y-axis represent the size of uniform, the case with the aspect of

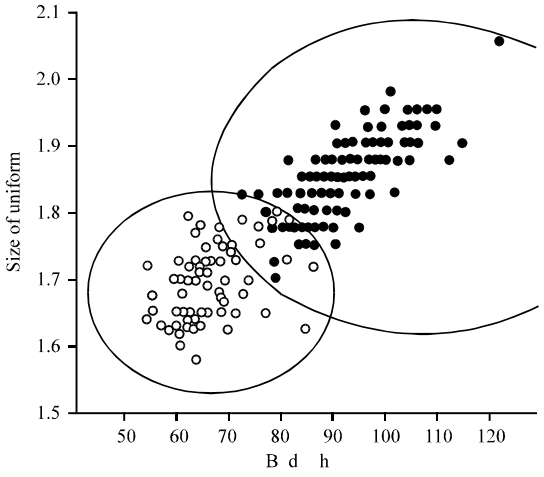

Fig. 1: Clustering the sizes of uniform

various height, weight, short legged, small waist are taken for classification and clustered as even sizes, odd sizes, plus sizes, petite and so on. The database contained >100 measurements for each of 3,000 women.

NEITHER IS RIGHT AND NEITHER IS WRONG

In similar, there is no specific technique which can be opted for all kind of issues caused from different data. To acquire quality clusters, the earlier said key issues have to be more focused on and the earlier mentioned said techniques has to be blended in the right ratio such that the quality clusters can be obtained.

CONCLUSION

Data analysis depends on what direction researchers approach it from, just like blind man and elephant. If a blind man approaches the elephant from the front and touches the trunk, he may assume that it is a snake-like creature. If another blind man touches the elephant's leg, he may assume that the elephant is more like a tree.

REFERENCES

Gupta, S.C. and V.K. Kapoor, 2010. Fundamentals of Mathematical Statistics. 1st Edn., Sultan Chand and Sons, New Delhi.

Halkidi, M., Y. Batistakis and M. Vazirgiannis, 2002. Cluster validity methods: Part I. ACM SIGMOD Record, 13: 40-45.

Jeyabalaraja, V. and E.T. Prabakaran, 2012. Study on software process metrics using data mining tool: A rough set theory approach. Int. J. Comput. Applic., 47: 1-5.

Marx, Z., I. Dagan, J.M. Buhmann and E. Shamir, 2002. Coupled clustering: A method for detecting structural correspondence. J. Machine Learn. Res., 3: 747-780.

Rai, P. and S. Singh, 2005. A survey of clustering techniques. Int. J. Comput. Appl., 7: 1-5.

Shtern, M. and V. Tzerpos, 2012. Clustering methodologies for software engineering. Adv. Software Eng. 10.1155/2012/792024

Sileshi, M. and B. Gamback, 2009. Evaluating clustering algorithms: Cluster quality and feature selection in content-based image clustering. WRI World Congress Comput. Sci. Inform. Eng., 6: 435-441.

Strehl, A. and J. Ghosh, 2002. Cluster ensembles: A knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res., 3: 583-617.

Yang, B., X. Zheng and P. Guo, 2006. Software metrics data clustering for quality prediction. Proceedings of the Part II International Conference on Intelligent Computing, August 16-19, 2006, Kunming, China, pp: 959-964.