Link based Cluster Ensemble Framework - Clustering Categorical Data for Internet Security Applications

S.Sugantha
PG Scholar
Anna University: Regional Centre
Thiruchirappalli

C. Ramasamy @ Sankar Ram
Assistant Professor
Anna University: Regional Centre
Thiruchirappalli

ABSTRACT

In recent years, an increasing number of security threats have brought serious risks to the internet. Internet security is needed for providing protection from internet related threats whose are threatening the availability of the internet, and the privacy of its users. One best solution for providing internet security is to use antivirus software product and it uses signature based detection method. Malware attacks and phishing websites (fake websites) are two major security threats. So we need an efficient method for automatically categorizing those threats for signature based detection .In this paper we propose a categorization system for profiling signatures to improve the anomaly detection process more efficiently. A categorization system that uses a link based cluster ensemble for automatically categorizing security threats. Cluster ensemble aggregates different clustering algorithms producing different solutions for grouping malware samples and phishing websites.

Index Terms - Hybrid Hierarchical Clustering Algorithm (HHCA), Link Based Cluster Ensemble (LBCE), Malware categorization, Phishing websites, Weighted K - Medoids Algorithm (WKMA)

1. INTRODUCTION

1.1 Malware categorization

Malware is the one of the major internet security threat. Currently, Antivirus (AV) software product is used for providing protrude signature profile for detecting malware. Modern malware is very complex and many variants of the same virus with different abilities appear every day which makes the detection process more difficult. For many years, malware categorizations have been done by human analysts such as looking up description libraries, and searching sample collections. The manual analysis is time consuming and subjective for handling huge data. An automatic categorization system is required for making malware detection more efficient.

1.2 Phishing Websites Categorization

Phishing is a fraudulent attempt to get personal information such as bank information, employment details, and online shopping account passwords and so on from victims. Phishing websites are designed to fool recipients into divulging personal financial data. Phishing problem is a hard problem because it is very easy for an attacker to create an exact replica of a good site. Phishing websites resembled as trustworthy websites to allure internet users for revealing their sensitive information. Security software products use blacklisting to filter these phishing websites against known websites. There is always a delay between website reporting and blacklist updating due to manual analysis. As the lifetimes of phishing websites are reduced to hours from days, this method might be ineffective.

Malware attack and fake websites are two different forms of Internet security threats and they are sharing several common properties. Both are driven for economic benefits and increasing rapidly .An effective method is needed for categorizing these threats and it will helpful for anomaly detection process. Though the phishing websites and the malware samples evolve constantly, most of their inherent structure is relatively stable. A family of malware samples typically exhibit similar behavior profiles [3]. Over the past few years, many clustering algorithms have been developed for automatic categorization of malwares and for phishing website detection and prevention [14]. Phishing websites are not isolated from their targets but have strong relationships with them [13], which can be used as clues to cluster them into families and generate the signature for detection.

The detection process is generally divided into two steps: feature extraction and categorization. In the first step, features such as Application Programming Interface (API) calls and instruction sequences are extracted to capture the characteristics of the file samples and term frequencies of the webpage content. These features can be extracted via static analysis and/or dynamic analysis. For categorization step, intelligent techniques are used to automatically categorize the file samples or the websites into different classes based on computational analysis of the feature representations. Different clustering algorithms and even multiple trials of the same algorithm may produce different results due to random initializations and stochastic learning methods [8].

Anomaly detection refers to the problem of finding patterns in data that do not conform to the expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains. Anomalies might be induced in the data for a variety of reasons, such as malicious activity, e.g., credit card fraud, cyber-intrusion, terrorist activity or breakdown of a system, but all of the reasons have a common characteristic that they are interesting to the analyst [15]. The real life relevance of anomalies is a key feature of anomaly detection.

1.3 Contribution of the Paper

In this paper, first we observe the phishing websites and malware samples represented in terms of term frequency of the webpage content and instruction frequency of the program. Then we develop a categorization system for grouping phishing websites or malware samples into a class that share some common characteristics using link based cluster ensemble. Cluster ensemble aggregates different clusters that are produced by different clustering algorithms.

To improve the clustering performance and conventional cluster ensembles, we develop Link - based Cluster Ensemble (LBCE) approach [10] for aggregating different base clustering solution. The result of the cluster ensemble is used as a signature profile in anomaly detection system. Anomaly detection system uses this signature profile for detecting fake websites and malware attacks to provide internet security. Our categorization system has the following features:

- **Feature representation:**
  Term frequency of the webpage content is used to represent websites, while instruction frequency is used for malware feature expression. These features well represent variants of phishing websites and malware families, respectively, and both can be efficiently extracted. We use a uniform framework which is based on clustering ensemble for both Internet security threats.
- **Well designed base clustering algorithms:**
  To handle the instruction frequency features, we use both HHCA and WKMA algorithms to generate base clustering.
- **A novel link based cluster ensemble scheme:**
  A new LBCE approach is more efficient than the former model. It is used to generate accurate and inexpensive measures. A link based similarity algorithm (LBSA) is used for this purpose.

2. RELATED WORK

2.1 Malwares and Phishing Websites Categorization

2.1.1 Feature Extraction and categorization

Features are the characteristics of the program under analysis. There are various three categories of feature extraction methods: dynamic, static, and hybrid. Dynamic analysis techniques observe the execution of the malware to derive features. Well known techniques include debugging and profiling. One advantage of dynamic feature extraction is that the environment or configuration dependent information has been resolved during the extraction, e.g. a variable whose value depends on the hardware, system configuration, or program input. One disadvantage of dynamic analysis is its limited coverage. Static analysis techniques analyze the malware without running it. Static analysis has the advantage that it can explore all possible execution paths in the malware. One disadvantage of static analysis is its inability to address certain situations due to undecidability. Hybrid analysis is an approach that combines static and dynamic analysis to gain the benefits of both.

Phishing website is a semantic attack which targets the user rather than the computer. Recently, many classification methods such as support vector machines and Naive Bayes have been used for anti-phishing. The most common methods used today for the detection and analysis of phishing web sites [16] are:

- Manual view and report services such as Phishtank.com
- Correlating links in known spam email to phishing sites
- Crawler classification of websites

Given an unknown webpage, Liu *et al.* [18], [19] proposed the following method for phishing detection and clustering: For detection ,the method first finds the associated webpages with the given page, then mines the features (such as links relationship, ranking relationship, webpage text similarity, and webpage layout similarity relationship) between the given webpage and its associated webpages, and finally applies DBSCAN (Density Based Spatial Clustering of Applications with Noise) clustering algorithm to decide if there is a cluster around the given webpage. If such cluster is found, the given webpage is then regarded as a phishing webpage; otherwise, it is identified as a legitimate webpage. For clustering, it first extracts the bag-of-word representation from the source of the websites and then principal component analysis (PCA) for feature selection, and, finally, uses certain clustering algorithms (such as *k*-means, DBSCAN) for detection. For example, the experiments were performed based on 8745 phishing webpages and 1000 legitimate webpages, while Layton *et al.* [19] evaluated their proposed methods based on a dataset containing 24403 websites.

Various classification approaches including association classifiers, support vector machines, and Naive Bayes have been applied in malware and phishing website detection. In particular, existing clustering methods usually apply a specific clustering method on a feature representation. Different clustering methods have their own advantages and limitations in malware detection. In our study, we use a link based cluster ensemble to aggregate the clustering solutions that are generated by both hierarchical and partitional clustering methods. Our ensemble framework is also able to incorporate the domain knowledge in the form of sample level constraints.

2.2 Cluster ensemble

Clustering ensemble obtains a single and better performing clustering solution from a number of different input clusters for a particular dataset [8]. Many approaches have been developed to solve ensemble clustering problems [10]. However, most of these methods are designed to combine partitional clustering methods, and few have combined both partitional and hierarchical clustering (HC) methods.

3. SYSTEM ARCHITECTURE

Fig.3.1 shows the architecture of categorization system and we briefly describe each component below:

- **Feature Extractor:**
  Term-frequency feature extractor:
  For phishing website categorization, we use the term frequency feature extractor to extract the terms from the webpages of the collected phishing websites, and then transform the data into term-frequency feature vectors and stored in the database.
  Instruction-frequency feature extractor:
  For malware categorization, we use the instruction frequency feature extractor to extract the function based instructions from the collected Portable Executable (PE) malware samples, convert the instructions to a group of 32-bit global IDs as the features of the data collection, and store these features in the signature database.
- **Base Clustering Algorithms:**
  The choice of base clustering algorithms is largely dependent on the underlying feature distributions. To deal with instruction frequency features, we use HHCA algorithm and WKMA algorithm to generate base clustering.
- **Link Cluster Ensemble:**
  In a link-based cluster ensemble framework: a cluster ensemble is created from M base clustering and generates the refined cluster association matrix from the ensemble using a link-based similarity algorithm. Finally clustering result is produced by a consensus function of the clustered partition.

4. BASE CLUSTERING

A cluster is a collection of phishing websites or malicious files that share some common traits between them and are "dissimilar" to the phishing websites or malware samples belonging to other clusters. Hierarchical and partitioning clustering are two common types of clustering methods, and each of them has its own traits [20]. The HC method can deal with irregular dataset more robustly, while partitioning clustering like KM is efficient and can produce tighter clusters especially if the clusters are of globular shape.

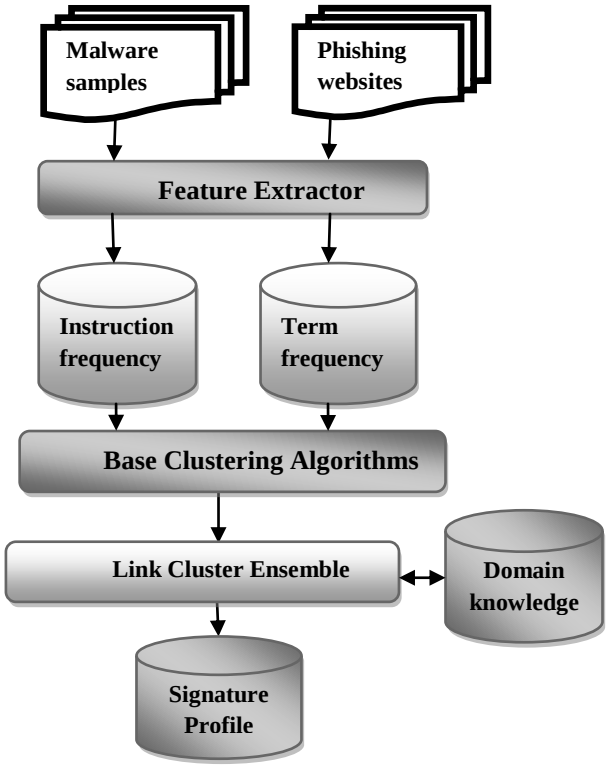

Fig.3.1 System Architecture

4.1 Hybrid Hierarchical Clustering Algorithm Algorithm(HHCA)

A Hybrid Hierarchical Clustering Algorithm (HHCA) combines hierarchical clustering and k-medoids algorithms to general base clustering. HHCA utilizes the agglomerative hierarchical clustering algorithm as the frame, starting with N singleton clusters, and merges the two nearest clusters until only one cluster remains. At each an iteration, HHCA adopts k-medoids algorithm to generate a partition. HHCA computes a cluster validity index at each of the iteration and generates the best number of clusters by comparing these indices. The outline of HHCA is described in Algorithm 1.

Input: The data set D
Output: The best K and data clusters
Set each sample as a singleton cluster;
For K N - 1 to 1 do
Merge two clusters with closest medoids;
Generate the new medoids of the merged clusters;
Run K-medoids to obtain a partition;
Calculate the validity index;
Compare and keep the best K and corresponding clusters
until now;
End
Return the best K and corresponding clusters.
Algorithm 1: Algorithm description of HHCA

We use Fukuyama-Sugeno index (FS) as the cluster validity index. FS evaluates the partition by exploiting the compactness within each cluster and the distances between the cluster representatives. It is defined as

$$FS = \sum_{i=1}^{N}\sum_{j=}^{nc}(\|x_i - v_j\|^2 A - \|v_j - v\|^2 A),$$

$$v = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Where, is the i[th] data point, v_j is the medoid of cluster C_j, v is the medoid of the whole data collection, is the membership value of the data of the cluster , m is the weighting exponent such that m ∈ , nc is the number of clusters and A is an 1×1 positive definite, symmetric matrix. It is clear that for compact and well-separated clusters, we expect small values for FS.

4.2 Weighted K - Medoid Algorithm (WKMA)

WKMA is used to generate base clustering on instruction sequences. WKMA dynamically assigns a weight to every feature for each malware family, which makes the clusters hiding in the subspaces and the common features of the same family can be easily generated. If a feature has a small variation within a cluster and large variations between the cluster and other clusters, then the feature can be viewed as an important feature for the cluster. Formally, denote the feature weight for cluster i as $\boldsymbol{W_i} = (w_{i1},\ldots,w_{id})$ where w_{ij} denotes the weight of the j[th] feature for cluster i and can be updated as follows:

$$w_{ij} = \begin{cases} \dfrac{\sum_{l=1}^{d} D_{il} - D_{ij} +}{(d-1)\sum_{l=1}^{d} D_{il} + \sum_{l}^{i}}, & \sum_{l=}^{d} > 0 \\ \dfrac{1}{d} & \text{Otherwise} \end{cases} \quad (1)$$

Where,

$$D_{ij} = \sum_{x_t \in C_i} w_{ij}(x_{tj} -^2, \quad = \quad \sum_{x_t \in C_i} w_{ij}(x_{tj} -^2$$

C_i is the i[th] cluster, and m_{ij} is the j[th] feature of the medoid for C_i. Note that $\sum| = 1$. Using the feature weight vector, we can compute the weighted distance between data points. The weighted distance is then used for computing the medoids and for assigning points into clusters. The algorithm procedure for WKMA is described in Algorithm 2.

Input: N points in d-dimensional space, number of clusters k
Output: k clusters and the corresponding weight vector
Randomly choose k cluster medoids;
Set initial weights to be ;
Repeat
Assign each point to the nearest cluster;
Update the cluster medoids;
Update the weight vector using Eq. (1);
Calculate the validity index;
Until the weight vectors and the medoids do not change;
Algorithm 2: Algorithm description of WKMA

5. LINK BASED CLUSTER ENSEMBLE (LBCE) FRAMEWORK

A link-based algorithm has been used to generate such measures in an accurate, inexpensive manner. The LBCE methodology is illustrated in Fig. 5.1. It includes three major steps of: 1. creating base clustering to form a cluster ensemble (π), 2. Generating a refined cluster-association matrix (RM) using a link-based similarity algorithm, and 3. Producing the final data partition (π*) by exploiting the spectral graph partitioning technique as a consensus function.

5.1 Cluster Ensemble

Let X = $\{x_1,\ldots,x_N\}$ be a set of N data points and π = $\{\pi_1,\ldots,\pi_M\}$ be a cluster ensemble with M base clusterings, each of which is referred to as an ensemble member. Each base clustering returns a set of clusters π = , …, such that =X .Where, k_i is the number of clusters in the i^{th} clustering. For each x X, C(x) denotes the cluster label to which the data point x belongs. In the i^{th} clustering, C(x) ="j" (or "C_{ij}") if x C_{ij}. The problem is to find a new partition of a data set X that summarizes the information from the cluster ensemble . Fig. 5.2 shows the general framework of cluster ensembles.

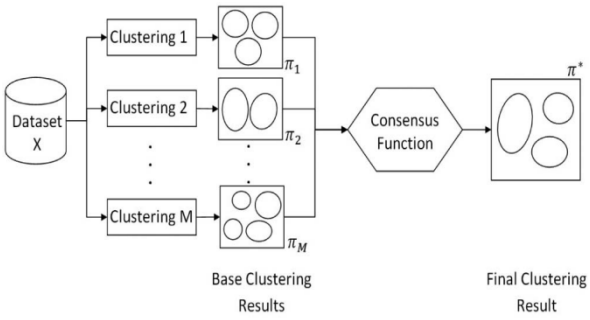

Fig.5.2 Process of cluster ensemble: It first applies multiple base clusterings to a data set X to obtain diverse clustering decisions (1 . . . M). Then, these solutions are combined to establish the final clustering result ($\pi_1,\ldots,\pi_M$) using a consensus function.

5.2 Refined Matrix (RM)

For each clustering π $_t$, t =1 . . . M and their corresponding clusters . . . $_t$ (where k_t is the number of clusters in the clustering π $_t$), the association degree RM (x_i,*cl*) [0,10] that data point x_i X has with each cluster *cl* . . . is estimated as follows:

$$\mathbf{RM}\ (xi, cl) = \begin{cases} 1, & if\ cl = \\ sim(cl, C_*^t(x_i)), & Other \end{cases} \quad (2)$$

Where, is a cluster label (corresponding to a particular cluster of the clustering π $_t$) to which data point x_i belongs. In addition, sim(C_x,C_y) denotes the similarity C_x,C_y, which can be discovered using the following link based algorithm.

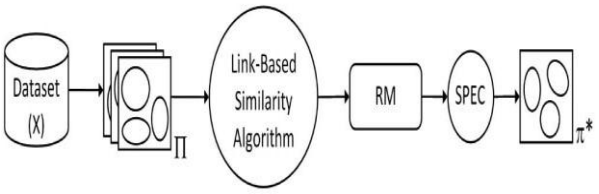

Fig.5.1 LBCE framework: 1. Cluster ensemble is created from M base clusterings, 2. Generate refined cluster association matrix from the ensemble using a link-based similarity algorithm 3. Final clustering result (π*) is produced by a consensus function of the spectral graph partitioning.

5.2.1 *Link Based Similarity Algorithm (LBSA)*

Given a cluster ensemble π of a set of data points X, a weighted graph G = (V,W) can be constructed, where V is the set of vertices each representing a cluster and W is a set of weighted edges between clusters. Formally, the weight assigned to the edge w_{xy} W, that connects clusters C_x, Cy V, is estimated by the proportion of their overlapping members.

Where,

$$\mathbf{w_{xy}} = \frac{|L_x}{|L_x}, \quad (3)$$

L_z denotes the set of data points belonging to cluster C_z V. Formally, a vertex C_k V is a common neighbor (sometimes called "triple," which is short for"center of the connected triple") of vertices C_x, C_y V , provided that w_{xk} , w_{yk} W. The weighted triple quality (WTQ) measure of clusters C_x, C_y V with respect to each triple C_k V is estimated by

$$\mathbf{WTQ_{xy}^k} \quad (4)$$

Here, W_k is defined as W_k = $\sum_{\forall t \in}$, where $N_k \subset$ V denotes the set of clusters that is directly linked to the cluster C_k, such that C_t N_k,w_{tk} W. The accumulative WTQ score from all triples (1 . . . q) between clusters C_x and C_y can be found as follows:

$$\mathbf{WTQ_{xy}} = \sum_{k=1}^{q} \mathbf{WXq} \quad (5)$$

The similarity between clusters C_x and C_y can be estimated by

$$\mathbf{sim(c_x,c_y)} = \frac{\mathbf{W}}{\mathbf{WT}}\ \mathbf{X}\ \mathbf{DC} \quad (6)$$

Where, WTQ_{max} is the maximum WTQ_{pq} value of any two clusters Cp,Cq V and DC [0,1] is a constant decay factor (i.e., confidence level of accepting two nonidentical clusters as being similar). With this link-based similarity metric, sim(C_x ,C_y) [0,1]with sim(C_x, C_y) = 1, C_x, C_y V.

Input G = (V, W), a weighted graph, where C_x, C_y V;
$N_k \subset$ V, a set of adjacent neighbors of C_k V;
W_k= $\sum_{\forall t \in}$;
WTQ_{xy}, the WTQ measure of C_x {and} C_y;
WTQxy 0;

For each
c N_x;
If c N_y;
WTQ_{xy} WTQ_{xy} + ;

Return WTQ_{xy} ;

5.3 Consensus Function

Given an RM representing associations between N data points and P clusters in an ensemble π, a weighted graph G = (V ,W) can be constructed, where V = $V^X \cup V^C$ is a set of vertices representing both data points V^X and clusters V^C. W denotes a set of weighted edges that can be defined as:

- w_{ij} W when vertices v_i, v_j V^X.
- w_{ij} W when vertices v_i, v_j V^C.
- Otherwise, w_{ij} = RM (v_i,v_j) when vertices v_i V^X and v_j V^C. Note that the graph G is bidirectional such that wij is equivalent to w_{ji}.

SPEC applies k-means to these embedded points in order to acquire the final clustering result.

6. CONCLUSION

In this paper, we have developed categorization system which can be applied for phishing website categorization and malware samples into groups that share some common traits by a link based cluster ensemble approach of different clustering solutions are generated using different clustering methods. The prominent future work includes an extensive study regarding the behavior of other link based similarity measures.

7. REFERENCES

[1] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proc. APWG eCrime Res. Summit*, 2007.

[2] J. Azimi and X. Fern, "Adaptive cluster ensemble selection," in *Proc. 21s Int. Joint Conf. Artif. Intell.* San Francisco, CA, 2009.

[3] M. Bailey, J. Oberheide, J Andersen, Z. M. Mao, F. Jahanian, and J. Nazario, "Automated classification and analysis of internet malware," in *Recent Advances in Intrusion Detection*, (Lecture Notes in Computer Science vol. 4637). New York: Springer, 2007.

[4] S. Basu, I. Davidson, and K. L. Wagstaff, Eds., "Constrained Clustering: Advances in algorithms, Theory, and Applications," Boca Raton, FL: CRC Press, 2008.

[5] U. Bayer, P. M. Comparetti, C. Hlauschek, C. Kruegel, and E. Kirda, "Scalable, behavior-based malware clustering," in *Proc. 16th Annu. Netw. Distributed Secur. Symp.* , 2009.

[6] C. Herley and D. Florencio, "A profitless endeavor: Phishing as tragedy of the commons," in *Proc. New Secur. Paradigms Workshop*, 2008.

[7] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C.Mitchell, "Clientside defense against web-based identity theft," in *Proc. 11th Annu. Network Distrib. Syst. Secur. Symp.* , 2004.

[8] Strehl· and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, Mar. 2003.

[9] Y. Ye, T. Li, Y. Chen, and Q. Jiang, "Automatic malware categorization using cluster ensemble," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010.

[10] Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price ," A Link Based Cluster Ensemble Approach for Categorical Data Clustering," vol. 24, no. 3, march 2012.

[11] Rui Xu and DonaldWunsch." Survey of clustering algorithms,"*IEEE transactions on neural networks*, 16, May 2005.

[12] Y. Zhang, A. Fu, C. Cai, and P.Heng,"Clustering Categorical Data," Proc. Int'l Conf. Data Eng. (ICDE), 2000.

[13] G. Liu, B. Qiu, and L. Wenyin, "Automatic detection of phishing target from phishing webpage", in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010.

[14] R. Dazeley, J. L. Yearwood, B. H. Kang, and A. V. Kelarev, "Consensus clustering and supervised classification for profiling phishing emails in internet commerce security," in *Knowledge Management and Acquisition for Smart Systems and Service*, New York, Springer-Verlag, 2010.

[15] Alexander Y. Liu and Dung N. Lam, "Using Consensus Clustering for Multi-view Anomaly Detection," 2012.

[16] Joshua S. White, Jeanna N. Matthews and John L. Stacy "A Method for the Automated Detection of Phishing Websites through both Site Characteristics and Image Analysis", 2012.

[17] Maher Aburrous, M. A. Hossain, Keshav Dahal, Fadi Thabtah, "Predicting Phishing Websites Using Classification Mining Techniques with Experimental Case Studies, "Information Technology: New Generations, Third International Conference on, pp. 176-181, 2010 Seventh International Conference on Information Technology, 2010.

[18] G. Liu, B. Qiu, and L. Wenyin, "Automatic detection of phishing target from phishing webpage", in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010.

[19] R. Layton and P. Watters, "Determining provenance in phishing websites using automated conceptual analysis", in *Proc. eCrime Res. Summit*, 2009.

[20] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, May 2005.