Applied Mathematics & Information Sciences
*An International Journal*

http://dx.doi.org/10.12785/amis/090659

# Data Analysis of Medical Records in Veterinary Hospital Using Clustering Method and Association Rule

*Tsung-Chih Hsiao* [1], *Shu-Ling Shieh* [2,*], *Tzer-Long Chen* [3], *Chia-Hui Liu* [4] *and Ying-Chi Yeh* [5]

[1] College of Computer Science and Technology, Huaqiao University, China.
[2] Department of Information Networking and System Administration, Ling Tung University, Taiwan.
[3] Department of Creative Product Design, Ling Tung University, Taiwan.
[4] Department of Digital Literature and Arts, St. John's University, Taiwan.
[5] The Graduate Institute of Applied Information Technology, Ling Tung University, Taiwan.

Received: 18 Mar. 2015, Revised: 16 May 2015, Accepted: 17 May 2015
Published online: 1 Nov. 2015

**Abstract:** Since pets have been promoted from working livestock to family companions, pet industry is also progressing with the development of nowadays? society. To reduce the negligence of the diagnostic process, the correlations between the symptoms and treatments of sick pets are collected. Medical records of a veterinary hospital are used as the input dataset. The clustering algorithms and association rules use two methods to investigate the practice. The result indicates that urology, dermatology or internal medicines are highly correlated with possible symptoms. The conclusion could assist doctors to find the association rules between diseases and symptoms of pets. In this paper, the medical records of pets are analyzed using the data mining method of Clustering Algorithm based on Histogram Threshold (HTCA). HTCA is composed of hierarchical clustering method and Otsu?s method. Two steps are applied in the experiments. First, we find the proper clustering by applying HTCA clustering method, and then apply the association rule to find the impact of divisions of property. Experiment shows that, the attributes of impacting factors and to efficiently find medical records in veterinary hospitals.

**Keywords:** Data Mining; Clustering; Association Rule; Electronic Medical Records..

## 1 Introduction

The market of veterinary hospitals has been expended in the past decade in Taiwan because of the prevalence of keeping pets. However, since the medical market has been saturated because of the competition, various veterinary hospitals start to enhance service quality and information management. To face the enormous demands of customers, current medical managers expect to enhance veterinary hospitals? management efficiency by improving technology management and service quality. Since the development of technology, the traditional paper-based medical records have been replaced by electronic medical records. Therefore, various diagnostic data of pets have been completely recorded and more and more researchers adopt those animal electronic medical records to do data mining analysis. With electronic medical records, veterinarians could standardize medical records, and keep observing the health conditions of animals for the healthcare programs. The correlation ship between symptoms and diseases could be found from electronic medical records. Hence, the negligence in the diagnosis process could be reduced by referring the discovered correlations from the medical diagnoses. Since the information of customers and pets are numerous, it costs the problem that the past management could no longer handle the huge amount of demands. Therefore, veterinary hospitals have to manage such huge amount of demands, and store enormous information with new models. To enhance data correctness, reserve data permanently, and assist veterinarians in understanding pets? health conditions at any time, medical institutions are rapidly making the medical records into electronic data. To find interesting characteristics and phenomena from data is the major task in data mining technology (Fayyad, 1996). Analyzing medical records appropriately could discover the impacting factors in the diseases and the correlations among them. However, the diagnostic

* Corresponding author e-mail: ltcc63@teamail.ltu.edu.tw

data in general veterinary hospitals are hardly acquired in the consideration of privacy. Hence, pets? medical records are analyzed without involving in personal privacy in this paper. In this experiment, the medical records in veterinary hospitals was mined, and was utilized for exploring possible symptoms of such records highly related to Urology, Dermatology, or General Medicine, expecting to find out the disease correlations among common symptoms of pets. First, the HTCA clustering method is first used for searching for the clusters and association rules is further utilized for the attributes of the divisions to find the factors in the classification and organize the rules, from which animals? diagnoses could be further realized.

## 2 Related Work

### *2.1 Clustering Method*

Clustering is an important method in data mining. Clustering is usually used to find out the hidden information and patterns in a data set. Han (2001) defined data mining as acquiring knowledge or unknown valuable information from a large amount of data to assist in making decisions or predictions. Clustering method aims to meaningfully partition enormous data to cluster all data points in a data set based on the similarity. In regard to data clustering, many researchers combine clustering with classification to overcome shortcomings of different clustering algorithms to solve problems. Shieh (2008) applied grey relational analysis to stock investment in order to find out the investment portfolio of stocks. Clustering is usually used to find out the patterns, beneficial relations or hidden information in a data set [Agrawal, 1994; Hamerly, 2002; Hirschman, 2002; Otsu, 1979; Shieh, 2009; Shieh, 2008; Maulik, 2002; Liu, 2002; Perex, 2007; Shieh, 2012; Zhang, 2011]. Han (2001) divided clustering into partitioning method, hierarchical method, density-based method, and grid-based method, where partitioning method and hierarchical method were commonly utilized. Following the designated number of clusters, medoid-based or centroid-based partitioning method is used for clustering, and then partition data points according to the cluster center to generate an initial partition result, which is regenerated a new cluster center after calculations to gradually improve the partition results. Partitioning method normally uses distance as the evaluation indicator. It is commonly utilized because of the simple concept, fast operating time, and good expansibility. However, when a attribute appears great variance or raised face, the points are likely to approach the nearest center that they could not be effectively and accurately clustered; K-mean and PAM (Partitioning Around Medoids) are the current representatives.

### *2.2 Hierarchical Clustering*

Johnson (1967) divided hierarchical clustering into agglomerative method and divisive method. Hierarchical clustering based on Euclidean distance is used to measure the similarity, presents favorable effects on dealing with simple and distinct convex figures. Agglomerative Method is generally utilized, where various data points are regarded as a single cluster, which is gradually merged, according to the distance, until the remaining number of clusters conforms to the pre-set value. On the contrary, divisive method regards the entire data set as a cluster, which is partitioned into two smaller clusters each time till the remaining number of clusters conforms to the pre-set value. Divisive hierarchical clustering is applied in this paper. Murtagh (1983) organized six distance evaluation methods in agglomerative method, namely Single Link Method, Group Average Method (Average Link Method), Centroid Method, Median Method (Gower?s Method), Minimum Variance Method (Ward?s Method), and Complete Link Method. Different evaluation methods would reveal distinct agglomerative results.

### *2.3 K-means Clustering*

K-means Clustering (Anwiti, 2012) is a Centroid Divisive Clustering Algorithm which is commonly used in partition. K-means could be applied to various types of data, and the clustering efficiency is high because of the less computation time and complexity. K-means aims to cluster n data points into k clusters so that the data points in the clusters show the highest similarity of intra-cluster, while the data points in distinct clusters reveal the lowest similarity of inter-cluster. The algorithm processes for K-means Clustering are shown as below. Step 1: Give an initial k cluster centers. Step 2: Designate each data point to the cluster with the closest cluster centers. Step 3: Update the center of each cluster. Repeat Step 2 and Step 3 till the cluster centers no longer change. Boecker et al. repeatedly partitioned data with K-means till there was merely one data point in each cluster and rapidly structured the process as a tree. Chen(2002) integrated K-means Clustering and Hierarchical Clustering and applied the initial clustering results with Hierarchical Method to the initial center and the number of clusters of K-means so as to solve the unstable clustering results. Hamerly (2002) indicated to set the initial cluster center with two methods, namely Forgy and Random Partition. According to the number of clusters k, the former randomly selected k data points from the entire data set as the initial cluster center. The latter, on the other hand, randomly designated data points into a cluster whose cluster center was calculated to be the initial cluster center. Forgy Method revealed dispersed initial clusters, while the cluster center with Random Partition Method focused on the center of the data set. According to

Hamerly et al., setting the initial cluster center with Random Partition Method would achieve better results.

### *2.4 HTCA Algorithm*

Hierarchical Threshold Clustering Algorithm (HTCA) (Shieh, 2012) is a dividing hierarchical clustering algorithm based on histogram. There are two steps in the HTCA Clustering: (1) partition method, and (2) hierarchical algorithm. In partition method, the values of data point in the clustered data set have to be transferred so that they could be mapped into several fixed integral intervals. Otsu (1979) threshold is used for automatically searching for the threshold. All attributes of the data in clusters are looked for the best partitioning point with the same method in order to acquire the best partitioning vector L. Assuming two cluster centers presenting the following relation, nearest neighbor method is applied to distributing the data points in a space to the nearest cluster to complete former partitioning step. Based on the hierarchical tree algorithm steps, several partitioning steps are preceded till the pre-set number of clusters k is conformed.

$$L = \frac{(m_1 + m_2)}{2}$$

### *2.5 Association Rules*

Association Rules, also named Market Basket Analysis, was proposed by Agrawal (1994). Association Rules would judge the meaning of rules according to the minimum Support and Confidence threshold pre-set by the user, which could be divided into the steps of (1) generating all high-frequency item set satisfying with the minimum Support, and (2) deducting all Association Rules satisfying the minimum Confidence, according to the generated high-frequency item set. For example, the deals in a transactional database TID1, TID2, TID3, TID4 are shown in table 1, and the number of transaction times of each item is shown in table 2. Assuming the pre-set threshold 50%, presenting 50% of the four deals in the list, the number of transaction times larger than 2 is regarded as high-frequency item. In this case, D and E are removed at the stage, and merely A, B, and C are the first-order high-frequency items. Furthermore, the items A, B, A, C, and B, C are generated by combining such first-order high-frequency items. The database is re-scanned and the number of appearing times of each candidate item is calculated, table 3. After the calculation, merely A, B conforms to the threshold that A, B is the high-frequency item at the stage. Since A, B cannot be combined for the candidate item at the next stage, it is therefore regarded as the high-frequency item of the database.

The acquisition steps with Apriori Algorithm are shown as below.

**Table 1:** Transactional database

| Transaction No. | Item |
|---|---|
| TID1 | {A,B,C} |
| TID2 | {A,B} |
| TID3 | {B,D} |
| TID4 | {C,E} |

**Table 2:** Number of appearing times of items.

| Deal No. | Number of appearing time |
|---|---|
| A | 2 |
| B | 3 |
| C | 2 |
| D | 1 |
| E | 1 |

**Table 3:** Number of appearing times of items.

| Deal No. | Number of appearing time |
|---|---|
| AB | 2 |
| AC | 1 |
| BC | 1 |

1.Find out the high-frequency (k-1)-item set, k¿1, and stop when it is an empty set ∅.
2.Find out the high-frequency (k-1)-item set with any two k-2 items from (1) to compose a k-item set.
3.Judge whether the sub-set with (k-1)-item set in k-item set found in (2) appears in (1). If so, the k-item set is remained; otherwise, it is removed.
4.Check whether the k-item set acquired from (3) satisfies the minimum Support. If so, it becomes the high-frequency k-item set; otherwise, it is removed.
5.Calculate Association Rules formed with high-frequency k-item set. When the minimum Confidence is satisfied, Association Rules is supported.

Association Rules is also widely applied to medicine. Kamal (1997) integrated Association Rules and Classification into medical diagnoses to predict other possible test results with several single-item medical test results.

## 3 Applying HTCA to Data Analysis of Medical Records in the Veterinary Hospitals

Data mining technology are applied in this paper to analyze the medical records in veterinary hospitals. First, HTCA clustering algorithm is used to find a number of key factors in the diseases. Then we apply the association rules to find the attributes of correlations among them. Those data that are not adequately correlated are filtered, and the rest of data are mapped and stored in proper format. Secondly, we apply the Apriori algorithm to

analyze the corresponding relation among the attributes. The pets? medical dataset is an animal hospital in the central of Taiwan. The dataset has diagnosis tables, customer tables, illness tables, and symptom. There are more than 600 original records. Those improper, duplicated, irrelative records are filtered and the required fields are finally selected. After pre-processed have four features: symptom, gender, weight, and body temperature are selected as data attributes. To achieve a better cluster result and to avoid the negative effects produced by noised and outliers, the dataset was pre-processed using data cleaning normalization. Since the experimental dataset is collected and categorized from the customer tables and illness tables, we try to summarize them into 3 kinds of illness. They are urinary, dermatology and internal medicine. After these data are rearranged and categorized, those duplicated records and records with null data are eliminated.

### 3.1 Normalization

In this paper, HTCA clustering algorithm is applied to judge the clusters result. Although the values of input data may be arbitrary values, large difference of values still may derive unpredictable results. To avoid the problem, a normalization pre-processing of the input data is required. The input data values are normalized into the range in [0, 1]. There are three well-known normalization methods: Min-max Normalization, Z-score Normalization and Normalization by Decimal Scaling. The method of Normalization by Decimal Scaling is used as following formula. The number of decimal points moved depends on the maximum absolute value of A. A value, $v_i$ , of A is normalized to $v_i'$ by computing $v_i' = \frac{v_i}{10^j}$ where j is the smallest integer such that $\max(|v_i'|) < 1$. Where $v_i$ is the value before the normalization process and $v_i'$ is the value obtained from the process according to $v_i'$ . This normalization method will map the given values into the range in [0,1].

### 3.2 HTCA clustering algorithm

We apply the HTCA clustering algorithm to find the clusters among the given data. The main process starts from the dividing, and it maps the entry points in the dataset into a finite range of integers. Then by the means of weighting divergence, we add MSE evaluating indices to calculate the results in each time we choose dataset. The cluster with looser data points spread would be choose first as the candidate to be split into two clusters. This splitting process is repeated until the number of sub-datasets equals to the default cluster number. We use the technology of clustering to find k groups from the diagnostic data. The result might be a reference to pets? illness. The input arguments of HTCA is a dataset X, where each record in X , $X_i$, i =1,2,? n. the number n is the number of total records, and k is the clustering number while the data resolution number is r. The output C* is the sets of Ci, where i = 2,3,?,k.

The algorithm of HTCA:

$$Input: X = [x_1, x_2, ..., x_n], k, r : Output : C* = [C_1, C_2, ..., C_k].$$

#### *Step 1: Initialization*

The data in the dataset X are mapped into an integer range [1, r] by means of proper mapping algorithm. Let the original dataset X be the initial cluster $TC_1$ . Then let the data cluster dataset numbered t, t = 1. The total divided times is set to be 1. Then use the function Size( ) to get the total number of rows n. Each record in the dataset is with R columns.

$$Transform\ X\ to\ [1,r] \in Z;$$
$$TC_1 = X;$$
$$t = 1;$$
$$[n,R] = Size\ of\ X;$$

#### *Step 2: Find splitting Vector*

Assume that there exists data in the target cluster and then calculate the number of values in each dimension. This will generate a histogram H of the data. According to each dimension i, there is a histogram $H_i$? i = 1, 2,..., R . By means of Otsu?s bi-level threshold we will find a proper separating point $L_i$? i = 1, 2,..., R, in dimension i. After each proper point is found in its dimension, these value Li will be collected as a best separating vector L. Because we use MSE to select the target cluster $TC_t$, the target cluster will be an empty set.

$$For\ each\ R\ as\ i$$
$$H^i = Histogram\ count\ of\ TC_t^i\ to\ [1,r];$$
$$L^i = Otsu's\ bi-level\ thresholding\ of\ H^i;$$
$$End\ of\ R$$

#### *Step 3: Assign Data Vector to Clusters*

The data in the target cluster are separated into two sub-clusters by assuming the primary center in the target cluster be the center $vm_1$ of the first sub cluster and the center of the second sub cluster, $vm_2$ , be the result of twice of the best separating point $L$ subtracting the center of the first sub cluster $vm_1$. This result is the mirroring mapping based on the best separating point $L$. Then calculate the divergence of the data fields for each data point in the target cluster $TC_t$ through choosing the

shortest distance using the divergence-weighted method to the two new centers. The new centers $vm_1$ and $vm_2$ will be also multiplied by the same weights, which are the weights of divergence $SL$ in each dimension, while separating the data in the target cluster into sub-clusters $TC_{2t}$ and $TC_{2t+1}$ , and computing the distances. These weighted data points are gathered as a vector $x'$ ,and the centers are named $vm_1{}'$ and $vm_2{}'$ respectively. When the distances are calculated, the vectors, $\| x' - vm_1{}' \|$ and $\| x' - vm_2{}' \|$, are applied into the calculation. Then we find the sub-cluster with the largest MSE value as the next target to be separated, where $i = 1,2,...n$ from all target clusters $TC_t$. After each time of the execution, the count of separated times is increased by 1.

$$
\begin{aligned}
&vm_1 = Centrold\,of\,TC_t;\\
&vm_2 = 2 * L - vm_1;\\
&For\,each\,R\,as\,i\\
&\quad SL^i = Diversity\,of\,TC_t^i;\\
&End\,for R\\
&[TC_{2t}, TC_{2t+1}] = TC_t\,which\,is\,Nearby\,vm_1, vm_2\,using SL\,as\,Diversity\,weighting:\\
&TC_t = Null;\\
&t = arg\,max\,MSE(TC);\\
&times = times + 1;
\end{aligned}
$$

### *Step 4: Termination Condition*

Let the non-empty cluster $TC$? be a terminal node of this binary tree. If $TC$? is less than the cluster number $k$ , and the number of separation is less than $k$?1, then go back to step 2 to continue the separating process. Otherwise, let $C$ be the terminal node of binary tree and then return it as the result.

$$
\begin{aligned}
&TC' = Terminal\,node\,of\,TC\\
&If\ |\ TC'\ | < k\,and\,times < (k-1)\,then\\
&\quad GotoStep2\\
&Else\\
&\quad C* = TC';\\
&\quad Return C*;\\
&End\,if
\end{aligned}
$$

### *3.3 Association Rule*

In this research, we apply the Association Rules to analyze the medical records in veterinary hospitals. Association Rules would judge the meaning of rules according to the minimum Support and Confidence threshold pre-set by the user, which could be divided into the steps of

1.generating all high-frequency item set satisfying with the minimum Support,

2.deducting all Association Rules satisfying the minimum Confidence, according to the generated high-frequency item set.

$$
\begin{aligned}
&L_1 = large1 - itemsets;\\
&for(k = 2; L_k \neq \varnothing; k++)\,do\,begin\\
&\quad C_k = apriori - gen(L_{k-1});\ //New candidates\\
&\quad forall\,transactions\,t \in D\,Do\,begin\\
&\quad\quad C_t = subset(C_k, t);\ Candidates\,contained\,in\,t\\
&\quad\quad for\,all\,candidates\,c \in C_t\,do\\
&\quad\quad\quad c.count++;\\
&\quad\quad end\\
&\quad\quad L_k = \{c \in C_k | c.count \geq min\,sup\}\\
&end\\
&Answer = U_k L_k;
\end{aligned}
$$

## 4 Experimental Results

In this experiment, we use MATLAB to simulate the experiment by the clustering methods of HTCA-2S, K-means. The platform of experiment is as follows. We use MATLAB R2010a as the analysis tool, SQL Server 2008R2 as the DBMS, and Microsoft Windows 7 Home edition (service pack1) as the OS. The software is running in a PC with Intel(R) Core(TM) i7-2600 CPU 3.40 GHz 3.40 GHz and 8 GB RAM. The source of pets? medical data is from a veterinary hospital in Taiwan. We re-arranged the data of diagnosis tables, customer tables, illness tables, symptom tables as the data source for pre-processing, and established effective data set. There are more than 600 original records, which were collected from pets? diagnostic data set. Each record consists of 56 attributes. The proportional value of each data classes is shown as table 4. The first class is the ratio of dermatology is 13.4%. The second class is the internal medical department is 78.6%. The third one is the urological department is 7.8%.

**Table 4:** Distribution table of dataset in each category.

| Class | Proportion |
|---|---|
| Dermatology | 13.4% |
| Internal medical department | 78.6% |
| Urological department | 7.8% |

The source of pets' medical data is from a veterinary hospital in the central of Taiwan. The rate of mis-clustering and its execution time in each group are showed in these data. We also analyzed the results of the clustering experiment. Each clustering method uses default constants in Table 5, where the K-means

algorithm is implemented using MATLAB. The resolution $r$ of HTCA algorithm is set by some proper values by its data fields respectively. The parameters of cluster number k are determined by the amount of categories in each dataset.

**Table 5:** Parameters used in the experiment.

| Method | Constants Set |
|---|---|
| HTCA | r is depend on the data set |
| K-means | MATLAB default |
| Apriori | Support=0.4, Confidence=0.2 |

Comparing the relation between the execution time and accuracy, the Accuracy Gain, which is the ratio of accuracy to the execution time, is defined as the following formula:

$$AG = \frac{(1-E)}{RT}$$

E is the error ratio of the results emulated by clusters to the mutual comparison among the real data sets. The value range of E is in $[0,1]$. The larger value of E implies the missing clusters distribution that is calculated by the method. And the small value of E represents the high accuracy of the clustering process. Let RT be the execution time spent by the clustering process, then AG is the gain of accuracy which is the ratio of improvement in a specific time unit. In the condition of small difference between two AG values generated by two methods, the method with higher AG will save more time. To assure the consistence of methods, the digital values in the data set of experiments are normalized into $[0,1]$. Those records with no data are moved from the clustering process.

(1) The results of HTCA clustering method In this experiment, we ran the HTCA and K-means on the pets? medical data. The error rate and execution time are shown in Table 6. We discover that the AG of result calculated by HTCA is 23% higher that results applying K-means. And the execution time of both experiments are almost the same. Therefore, the HTCA will be the method with the highest AG.

**Table 6:** The Accuracy Gain table in pets? medical caring data.

| Method | Mean Error Rate(E) | Mean Run Time(RT) | Mean Run Time(RT) |
|---|---|---|---|
| HTCA | 0.269 | 0.21 | 3.477 |
| K-means | 0.494 | 0.20 | 2.528 |

(2) The result of Association Rule Analysis We use the default association rule function in SQL Server 2008 to analyze the data set according to the pre-defined parameters. The result using Apriori algorithm is as following table.

In table 7, we may conclude that eye hyperemia may be a symptom of urological disease. It may me the symptom of chlamydia of cats, because Chlamydia may infect respiratory passages, digestion system and reproduction system. The symptom may include slight rhinitis, sneezing, conjunctivitis and tearing.

**Table 7:** Association rule in Urological Department.

| Probability | Importance | Rules |
|---|---|---|
| 1 | 0.91733 | Eye hyperemia(EYC) = true → category = Urological Department |

**Table 8:** association rules in dermatology.

| Probability | Importance | Rules |
|---|---|---|
| 1 | 0.71321 | Allergy (ALG) = true → category = dermatology |
| 1 | 0.71321 | Fleas (FLS) = true, Weight = 0 0.25 → category = dermatology |
| 1 | 0.71321 | jaundice(JAU) = true → category = dermatology |

In table 8, we may obtain the result that the symptoms of dermatology may include allergy, fleas and jaundice. Allergy may be caused by the drops of fleas or stung by fleas. Allergy may also be caused by jaundice.

**Table 9:** association rules of internal medicine.

| Probability | Importance | Rules |
|---|---|---|
| 1 | 0.03845 | Good appetite (GA) = true → category = internal medicine |
| 1 | 0.01848 | Hair change(HAC) = true → category = internal medicine |

In table 9, we may conclude that appetite and hair change may be the symptom of internal medicine department. The ill pets may be too fat or with the symptom of otitis externa. The external auditory meatus may be red or itch. There is yellow thick liquid in the external auditory meatus at first. When patient?s condition becomes worse, the thick liquid becomes black. This may cause puffiness and make the hair near ear fall off.

## 5 Conclusion

In this paper, we develop a methodology for applications of data clustering to efficiently find medical records in veterinary hospitals. HTCA Clustering is adopted to cluster the attributes of divisions for understanding the diagnosis situations of pets. The experimental results indicate that the symptoms affected by diseases could be mined through medical records so that we could understand the happening factor for treatment. The main process of our approach can be summarized as following. First, the accuracy of HTCA Clustering is higher than it with K-means Clustering. The result of this experiment shows that Clustering has a lower errors on the clustering of pets? medical record data sets than K-means clustering does. The performance improved execution time of the HTCA is much better than traditional method. Secondly, from the experimental analyses of Association Rules, the factors in red eyes might be classified into Urology,

possibly because of being infected with Chlamydia felis; allergy, flea, and choleplania are classified into Dermatology, possibly because of being infected with dog skin diseases; and, good appetite and hair change are classified into General Medicine as they are resulted from external otitis or obesity. The result indicates that the attributes of impacting factors and to efficiently find medical records in veterinary hospitals.

## Acknowledgement

This work was supported partially by the introduction of talents Huaqiao University Scientific Research Projects (Project No. 13BS412), Fujian Provin Engineering Technology Research Center of Green Communications and Intelligent Information Service (Project No. 2012H2002).

## References

[1] R. Agrawal & R. Srikant. Fast Algorithms for Mining Association Rules. IBM Research Report RJ9839, IBM Almaden Research Center, (1994).
[2] G. Hamerly & C. Elkan. Alternatives to the K-means Algorithm that Find Better Clustering. The Eleventh International Conference on Information and Knowledge Management, 600-607, (2002).
[3] J. Han & M. Kamber. Data Mining: Concepts and techniques. Morgan Kaufmann, (2001).
[4] L. Hirschman, J.C. Park, J. Tsujii, L. Wong & C.H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**, 1553-1561, (2002).
[5] N. Otsu, A Threshold Selection Method from Gray-level Histograms. IEEE Transactions on Systems, Man and Cybernetics, **9(1)**, 62-66, (1979).
[6] S. L. Shieh, I. E. Liao, K. F. Hwang & H. Y. Chen. An Efficient Initialization Scheme for SOM Algorithm Based on Reference Point and Filter. IEICE Transactions on Information and System, E92-D(3), 422-432, (2009).
[7] S. L. Shieh & I. E. Liao. A New Clustering Validity Index for Cluster Analysis based on a Two-Level SOM. *IEICE Transactions on Information and System*, **E92-D(9)**, 1668-1674, (2009).
[8] S. Shieh, K. Huang, C. Jane & D. Jheng. A New Grey Relation Analysis Applied to the Assert Allocation of Stock Portfolio. *International Journal of Computational Cognition*, **6(3)**, 6-12, (2008).
[9] S. Shieh, T. Lin & Y. Szu. An Efficient Clustering Algorithm Based on Histogram Threshold. *Lecture Notes in Artificial Intelligence*, **7197**, 32-39, (2012).
[10] U. Maulik & S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions Pattern Analysis and Machine Intelligence*, **24(12)**, 1650-1654, (2002).
[11] C. C. Liu, G. D. Chen, C. Y. Wang & C. F. Lu. Student Performance Assessment Using Bayesian Network and Web Portfolios. *Journal of Educational Computing Research*, **27(4)**, 437-469, (2002).
[12] H.B. Perex & F.G. Nocetti. Fault classification based upon self organizing feature maps and dynamic principal component analysis for inertial sensor drift. *International Journal of Innovative Computing, Information and Control*, **3(2)**, 257-276, (2007).
[13] S. L. Shieh & I. E. Liao. A new approach for data clustering and visualization using self-organizing maps. *Expert Systems with Applications*, **39(15)**, 11924-11933, (2012).
[14] U. M. Fayyad. Advances in Knowledge Discovery and Data Mining. *AAAI Press**, (1996).
[15] X. Zhang, J. Liu, Y. Du & T. Lv. A Novel Clustering Method on Time Series Data. *Expert Systems with Applications*, **38(9)**, 11891-11900, (2011).
[16] S. Johnson, Hierarchical Clustering Schemes. Psychometrika, **32(3)**, 241-254, (1967).
[17] F. Murtagh. A Survey of Recent Advances in Hierarchical Clustering Algorithms. Oxford University Press, (1983).
[18] J. Anwiti, R. Anand & B. Rupali. An Efficient K-Means Algorithm to Cluster Large Data-set in Data Minig. International Journal of Advanced Research in Computer Science and Electronics Engineering,**1(3)**, 86-91, (2012).
[19] Kamal Ali, Stefanos Manganaris, and Ramakrishnan Srikant, Partial classification using association rules. *In Knowledge Discovery and Data Mining*, 115-118, (1997).

**Tsung-Chih Hsiao** received the Ph.D. in the Department of Computer Science and Engineering, National Chung Hsing University, Taiwan. He is currently an instructor in the College of Computer Science and Technology at Huaqiao University, China. Research fields include Information Security, Cryptography, and Network Security.

**Shu-Ling Shieh** received the MS degree in information management from National Yunlin University of Science & Technology, Taiwan, in 1996, and the PhD degree in the Department of Computer Science of National Chung Hsing University, Taiwan, in 2004 and 2010, respectively. She is currently an associate professor in the Department of Information Networking & System Administration of the Ling Ting University, Taiwan. Her research interests are in Data Mining, Neural Network, Self-organizing Map, and Information Visualization.

**Tzer-Long Chen** received the Ph.D. in the Department of Information Management, National Taiwan University, Taiwan. He is currently an assistant professor in the Department of Creative Product Design at Lingtung University, Taiwan. Research fields include Information Security, Cryptography, and Network Security.

**Chia-Hui Liu** received the B.S. degree from Dayeh University in 2002, the M.S. degree from National Chiayi University in 2004, and the PhD degree from National Taiwan University in 2011, both in Computer Science, Taiwan. She is currently an Assistant Professor of Digital Literature and Arts at St. John?s University, and doing research, i.e., Information Security, Multimedia Technology and Cryptography.

**Ying-Chi Yeh** received the MS degree in the Graduate Institute of Applied Information Technology, Ling Tung University, Taiwan, in 2014. His research interests are in Data Mining, Neural Network, and Self-organizing Map.