Expert Systems with Applications 36 (2009) 6966–6977

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Computational accounting in determining Chart of Accounts using nominal data analysis and concept of entropy

P.Y. Wang, Tommy W.S. Chow *, Chris W.F. Chiu

Department of Electronic Engineering, City University of Hong Kong, Hong Kong

ARTICLE INFO

Keywords:
Computational accounting
Nominal data
Entropy

ABSTRACT

A new application of nominal data analysis to computational accounting system is introduced. "*Chart of Account*" (COA), the structure of the Enterprise Requirement Planning (ERP) Accounting System is used as a nominal dataset enabling nominal feature selection technique be used to determine the close-optimal number of accounting segments. Five datasets are simulated with entropy measurement using the method of SUD. Self-Organizing map (SOM) is used to investigate the similarities among different segments, which proved to be a useful approach in cross-examining the COA structure. The obtained results show that they are promising from both computation and accounting perspectives.

1. Introduction

Data mining has practical significance in many areas such as bioinformatics, statistics, pattern recognition, machine learning because it is the principle of dealing with large amounts of data and picking out related information. It is increasingly used in extracting information from the enormous data sets generated by modern experimental and observational methods. For instance, in the example of cancer-causing gene selection, numerical feature selection scheme is used to determine the dominant genes (features) from a given huge dimensions of a dataset with over tenth of thousands of features. Computational accounting, which includes financial accounting, managerial accounting and planning and control, is also an important application.

An Enterprise Requirement Planning (ERP) Accounting System is necessary for all medium and large size corporate company. The ERP Accounting System is used to reduce laborious manual work that results in an increase of the integrality in accounting sense. The foundation of the system is called the "*Chart of Accounts*" (COA) that is a primary and structural block consisting of a series of segments in numerical digits. In accounting principal, at lease two lines are used to record one transaction. One line represents a credit item, whose name is replaced by a code, and another line represents a debit item; the two lines must be equal for balancing. For example, in a case when a company purchases a computer server, Table 1, in which the code "300020" represents "Inventory", and the code "500010" represents "Liability", is used as the basic data input system. The numerical data shown in Table 1 is a three segments COA structure (Oracle General Ledge User Guide, 2003). This three-segment structure is the foundation of the COA structure nowadays.

The COA structure is the key for the ERP implementation; it is the first step for the implementation of an accounting system (Oracle General Ledge User Guide, 2003; Irv Chasen, 2006). Prior to the implementation of an accounting system, an accounting team is employed to design an COA structure for a company. It must be noted that the whole COA design is very time consuming, laborious, and costly. Also, a confirmed COA structure cannot be modified after commencement, because it records all the accounting information that includes profit and loss account, asset account, balance sheet account, and liability account. In practice, many corporate have had difficulty in carrying on using the same COA structure after a couple of years when the COA structure is found no longer in the most optimal way in representing the system effectively and efficiently. This is usually due to the time-variant nature of a modern corporate that their business patterns shift with time. But the ERP accounting system does not allow any minor COA restructuring after the system implementation. Thus, up to a certain stage, when a corporate finds the old COA structure started impeding the company operation, a complete revamp of the whole COA structure is required in order to maintain an effective daily operation. Developing the COA structure to date is purely based on accounting expertise. No one can assure to be able to determine a close-optimal COA structure or the optimal segment number, because the whole process is still man-dependent and heuristic. It is a lengthy and costly process. Although some corporate may try using the inappropriate COA structure despite the subsequent additional workload caused, a new COA structure must be designed because an inappropriate COA leads to an increase of operation cost and makes financial data difficult to analyze. In

* Corresponding author.
E-mail addresses: pywang@ee.cityu.edu.hk, eetchow@cityu.edu.hk (T.W.S. Chow).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.08.079

Table 1
Illustration of the balance between "credit" and "debit"

Account code	Debit	Credit
100.300020.00	10,000	
100.500010.00		10,000

practice, COA may require many different segments to represent an accounting transaction However, too many segments result in information broken down into discrete pieces, which makes the whole system very inefficient to construct an accounting line. In a medium size company that has about 2000 staff, and with an annual revenue and profit of about 0.25 billions US dollars, and the accounting lines are around 80,000 in each month, or about 2700 accounting lines per day. In view of the large number and complexity of accounting line to be handled daily, the process is very resources demanding and error prone (Enterprise Resource Planning, 2000).

In this paper, we describe a completely novel application of data mining using nominal feature selection technique to an accounting problem. Instead of spending months costly accounting expertise on finding an appropriate COA structure in a sense that it can best represent an accounting system in the least possible number of segments, the determination of COA in this paper is metamorphosed as a computational nominal feature selection problem. The contribution of this paper enables an expertise demanding accounting issue be handled computationally, which is the first of its kind.

Feature selection, whose objective is to build a simple and more comprehensible model, improving data mining performance, reduces features from a given feature set without performing transformation. It retains physical meaning of the selected features and provides clues for data collection or further analysis. There are numerous supervised feature selection methods (Chow & Huang, 2005; Huang & Chow, 2005). A supervised feature selection framework generally consists of two parts: a searching engine that is used to determine the promising feature subset candidates, and a criterion that is used to determine the best candidate. Nowadays, there are several types of searching engines. Heuristic searching, the most widely used approach, can deliver respectable results and can easily be implemented (Huang & Chow, 2004). However, when handling a huge nominal dataset computationally, unsupervised selection method is rare because in unsupervised learning, label, which is an important reference in supervised learning, is not given. Each unsupervised feature selection inherits characteristics of its employed clustering algorithm. Unsupervised feature selection schemes such as (Basak, De, & Pal, 1998; Dash, choi, Scheuermann, & Liu, 2002; Dy & Brodley, 2000; Mitra, Murthy, &

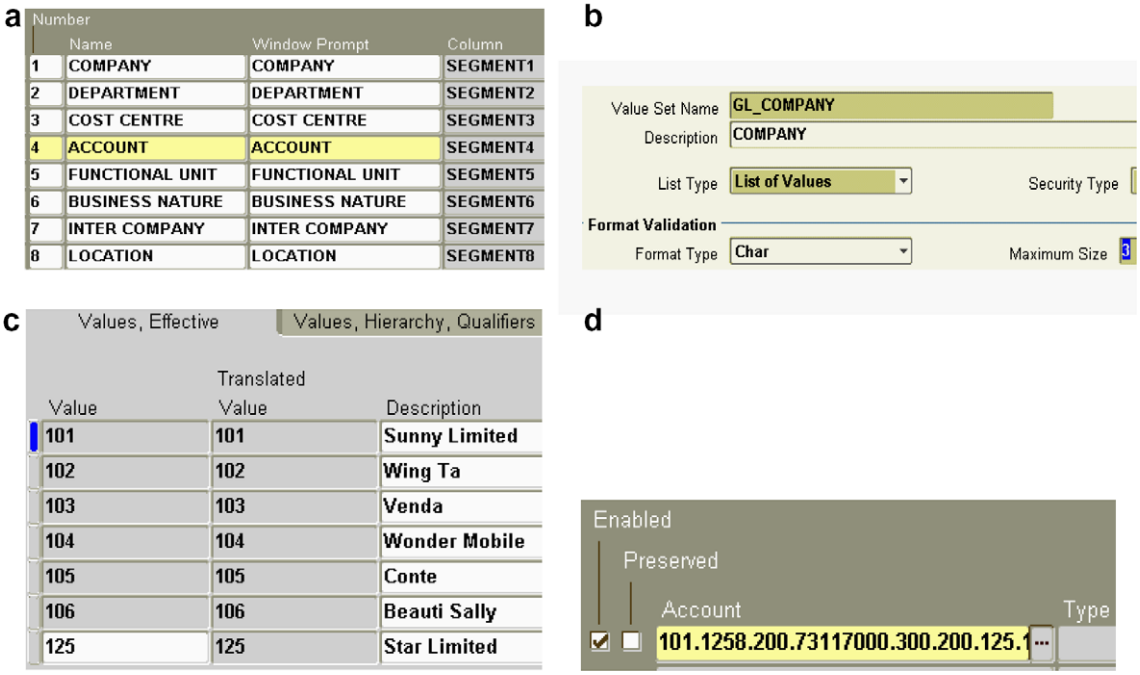

Fig. 1. A completed account code combination. (a) Setup COA structure, (b) setup the number of digits for each COA, (c) setup segment value and (d) setup segment code combination.

Table 2
Example of the existing COA

Segment number	1	2	3	4	6	7	8	9
Digits	3	3	3	8	3	3	3	3
Segment name	Company	Department	Cost centre	Account	Activity	Business nature	Mutual company	Location
Patterns	xxx	xxxx	xxx	xxxxxxx	xxx	xxx	xxx	xx
Code combination	101	1258	200	73117000	300	200	125	12
Code meaning	Sunny limited	IT	Infrastructure	Sales equipment	Hardware	Internal	Star limited	HK
Finalized COA structure								
Select/remove	Select	Select	Remove	Select	Remove	Select	Select	Select
Code meaning	Sunny limited	IT	NA	Sales equipment	NA	Internal	Star limited	HK

Pal, 2002; Shi & Suganthan, 2003) were reported. As the evaluation criteria of every feature selection scheme involves distance calculation, unsupervised feature selection schemes are unable to handle nominal data, which has no order information, such as color and brand name. Converting nominal data into binary data and inserting order into nominal data are common approaches to handle nominal data. Although this conversion makes the dataset filled with sparse features, which are usually treated as irrelevant features, it makes interpretation of a feature selection result possible.

In the design process, we first develop a loose COA structure that requires relatively little accounting expertise cost compared with the conventional way. In such a loose COA structure, it consists of many less essential or irrelevant segments that are redundancy features from the perspective of feature selection. Our objective is to use nominal feature selection approach to determine a close-optimal COA structure, or reducing the number of COA segments. In this feature selection process, we are able to determine the most effective features in a sense that they can best represent the accounting system. In this paper, we use the concept of entropy to calculate the relevance of all segments. SUD (Dash, Liu, & Yao, 1997), an unsupervised feature selection method, can handle nominal data without any transformation or conversion. It uses entropy similarity measurement to determine the importance of features with respect to the underlying clusters. By applying SUD, we can avoid the arbitrary order of the categories involved in nominal data. The application of the entropy concept to the COA feature selection problem enables us to directly decide the relevance of the segments. The segment with the least entropy is regarded as the most irrelevant.

The last section of this paper is to cross-examine the results from another perspective. As there are thousands of data points in each segment, we transform data points under every segment as high dimensional data, and segments are transformed as data points. We subsequently use Self-Organizing map (SOM) to examine the relationship between different segments. SOM, a well-known unsupervised visualization and clustering algorithm (Kohonen, 1995), has been widely used for numerous practical clustering and visualization applications (Chow, Rahman, & Wu, 2006; Chow & Wu, 2004; Wu & Chow, 2005). In (Magnusson et al., 2005; Marťın-del-Bŕıo & Serrano-Cinca, 1993), SOM is also employed for the analysis of finance related issues. A SOM output describes a mapping from a higher dimensional space to a lower

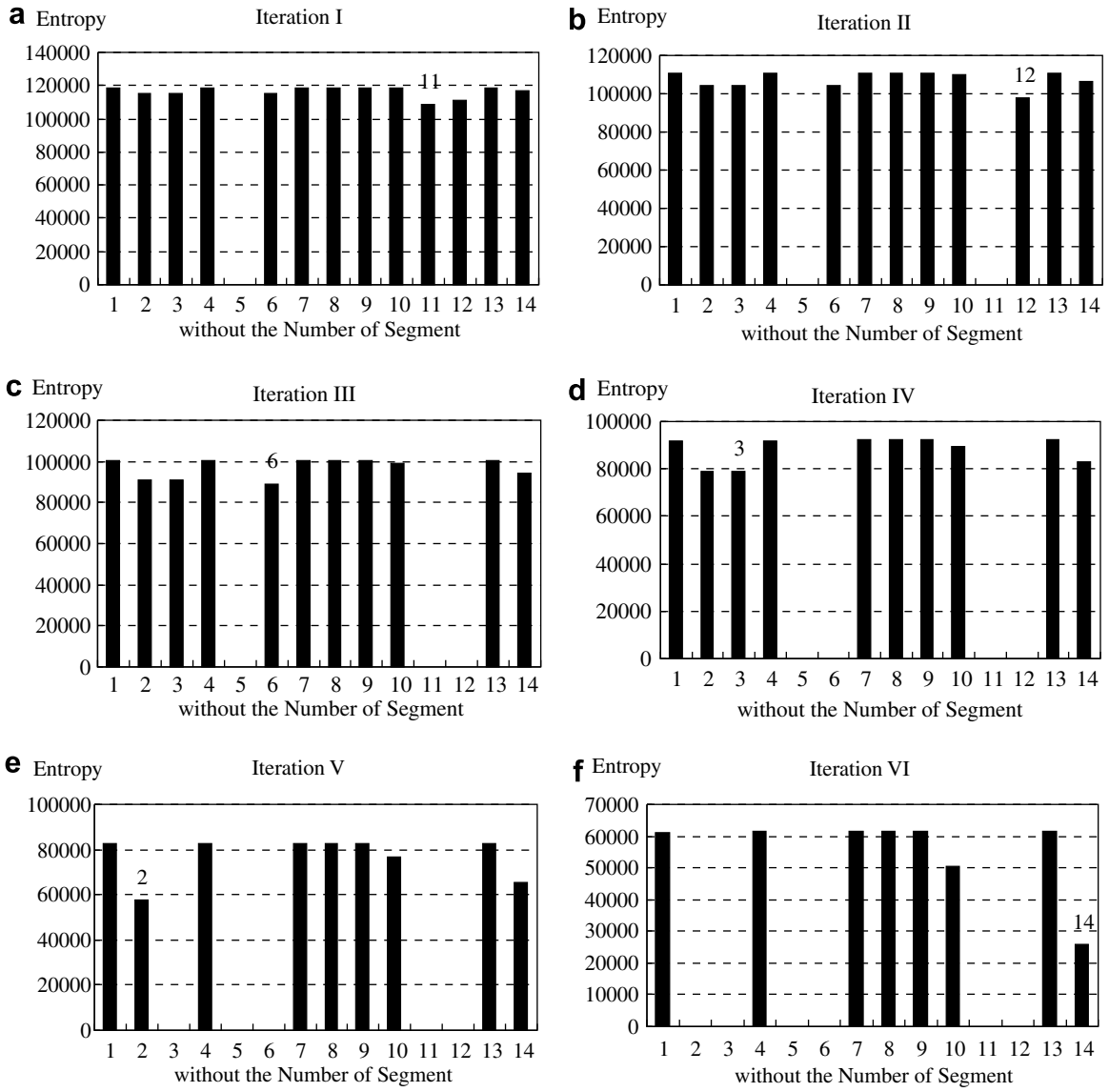

Fig. 2. Relationship of entropy when removing the number of segment for dataset "Capital Artist" iteratively.

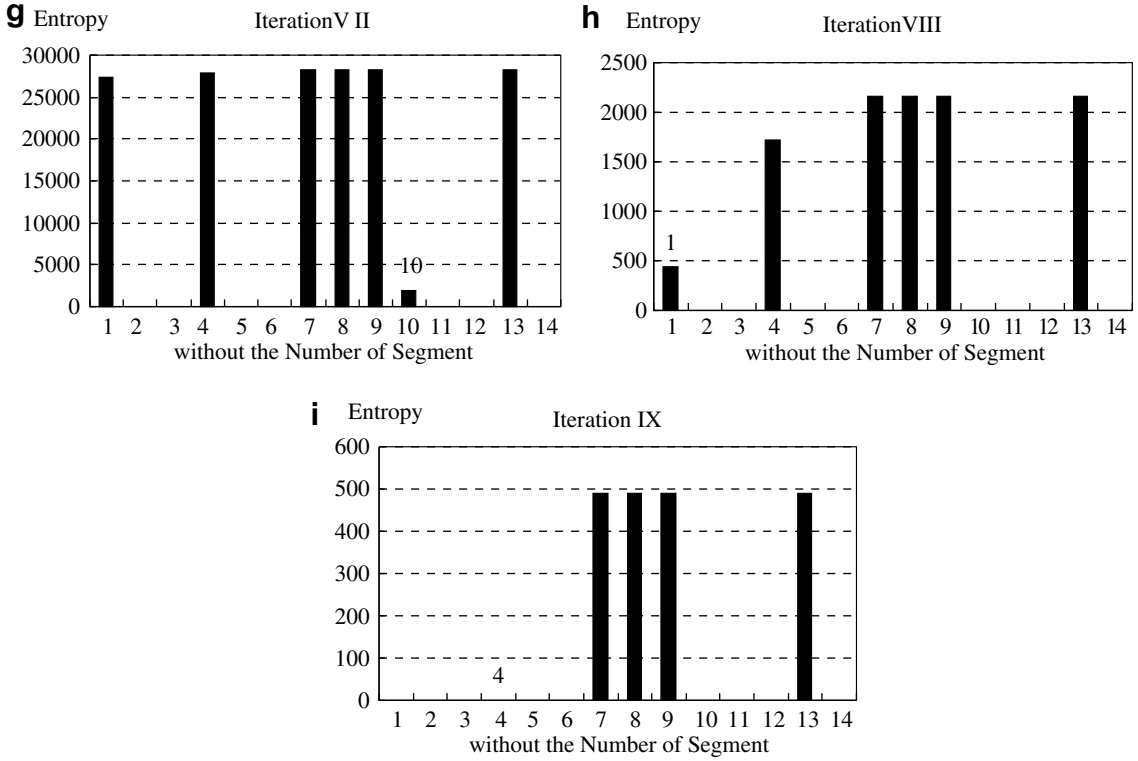

Fig. 2 (*continued*)

(usually in two-dimensional) dimensional space. The output map is useful because it preserves the data topological order making SOM map a kind of clustering that groups similar data together. In this study, similar segments in a COA structure are expected to exhibit certain extent of similar entropy measurements that are likely to be grouped together by SOM. The SOM output map can, thus, provides additional information in analyzing the construction of a COA structure.

This paper is organized as follows. Section 2 is the description of the detail COA structure physical meaning. Section 3 presents entropy measure for COA structure. Section 4 is the simulation results to illustrate why those segments are found to be irrelevant with the verification by the entropy theorem and SOM and the conclusion is shown in Section 5.

2. COA physical meaning

A typical COA structure is made up of a series of segments that consists of certain digits (Resource Planning, 2000). Each segment represents particular meaning. In an accounting system, a complete account code combination requires four steps to create. First, it is the design of a basic COA structure, such as the one demonstrated in Fig. 1 showing an eight-segment structure that ranged from "Company" to "Location" segments. Second, the number of digits of each COA segment is designed, for instance the "company" segment has three digits. We create a segment value, such as the "Sunny Limited" in the third step (Resource Planning, 2000). Finally, it is the generation of codes. The first two steps are the COA structure implantation. It cannot be changed once it is implemented. The following two steps are the continuous work after the system exists. Fig. 1a–d illustrates the flow of the four steps and a typical complete account code. Table 2 gives an example of the existing COA and there are altogether eight segments in one COA. The meaning from the 1st segment to the 4th segment is company, department, cost centre, account, while the meaning from the 6th to the 9th is activity, business nature, mutual company and location. For the 5th segment, it is the label of respective goods category. We put it in the middle as a reference during the implementation. The COA structure shown in Table 2 is a huge dataset that involves over 10,000 rows of data in a form of nominal data, consisting eight different features with the representation of segments here.

In this study, the COA structure consists of eight segments from "Company" to "Location". There are several digits to represent each segments and different digit has different meaning in terms of the detail contents in the specific segment. From Table 2, the accounting meaning of the code is read as *"Company Sunny Ltd, IT department, sells hardware, of infrastructure, to company Star Ltd, used by internal, in Hong Kong, Sunny Ltd. and Star Ltd. are mutual-companies*". By revealing the physical meaning of the COA structure, it is generated that digits in each segment stand not for the numerical meaning, but the nominal specification. There is no Euclidean distance among these digits and they cannot be presented in terms of geometric coordinates, which is one of the important differences between numerical data and nominal data. From this perspective, although COA structure is composed of segments with numerical digits, it should be regarded as the nominal data. Consider one COA code from the whole COA structure, which is separated by a set of segments. If one of the segments has distinct meaning, this segment is isolated to other segments. As one COA code has altogether eight segments now, and each segment has totally different meanings, the COA code must be well separated from segments. From the point of clustering, similar segments should be clustered together and dissimilar ones should be separated, irrelevant segments should not be influenced and get it trash. Relevant segments should be kept and they are useful on forming the COA code.

Nowadays, with more efficiency needed in the requirement, the structure of COA seems comparatively complicated just because

Table 3
Information of Five Dataset

	Capital artist	Investment	Property	Stock	Video film
Instances	870	1058	1194	110	1416
Features	14	14	14	14	14
Missing Values	0	0	0	0	0

COA still contains some redundant segments. COA structure nowadays which contains nine segments is derived from the past experience, which means that in the past there are more than nine segments in the COA structure. Up till now, no matter for the accounting expert or the application in the business, if we can take out those redundant segments before implementation, we will

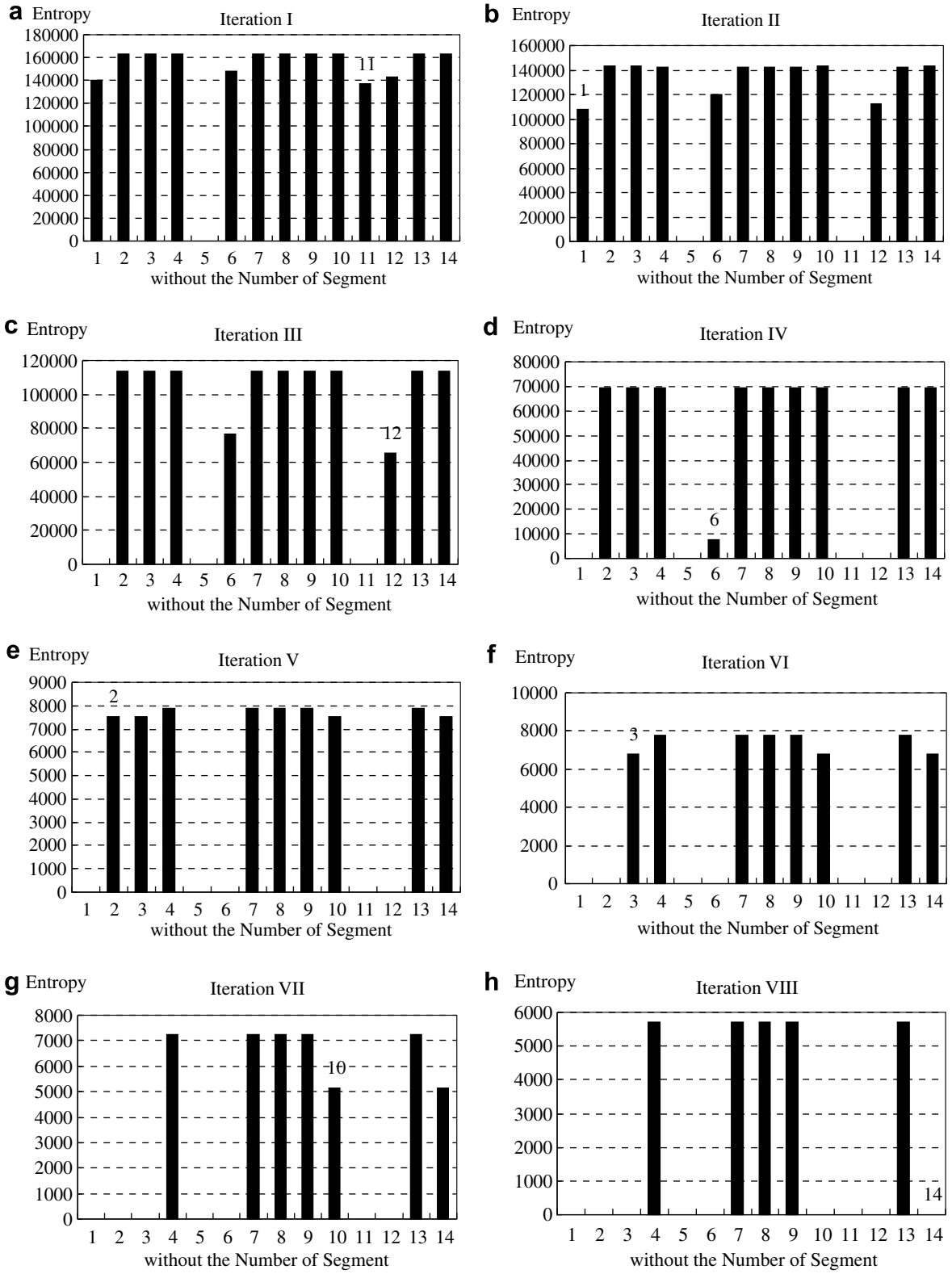

Fig. 3. Relationship of entropy when removing the number of segment for dataset "Investment" iteratively.

make COA simpler. Meanwhile, more profit will be made according to the new simpler COA structure.

3. Entropy measurement for COA structure

The entropy measurement of a COA structure is introduced in this section. In information theory, entropy is the measurement of the amount of information which is described as missing before reception. Sometimes it is referred to as Shannon entropy (Cover and Thomas, 1991). It is a more broad and general concept which finds applications in information theory. According to the introduction in the above section, data of COA structure can be regarded as nominal data. Entropy measure can be applicable to nominal data and does not need any class information for evaluating the entropy.

Dataset with m features and N nominal data instances, which are also referred as transactions in Yang, Guan, and You (2002); Yun, Chuang, and Chen (2001), is considered during the illustration of entropy concept. Nominal dataset consists of finite values in each feature and there is no order information among these values. For example, there is no order information in color, i.e., red is not closer to blue than green. The segment choosing method involves the application of entropy, which was originally used in classical thermodynamics. In statistical mechanics and Shannon theory (Cover and Thomas, 1991), entropy is defined as a measure of the number of ways in which the elementary particles of the system may be arranged under the given circumstances (Fast, 1962).

For a discrete distribution modeled by $X = \{x_1, x_2 \ldots x_N\}$, entropy measures the information which X is contained. Information means the uncertainty or the degree for a particular value of X being drawn. In a case that x is a value drawn from X, and the event $x = x_k$ occurs with probability p_k, the sum of the probabilities for $x = x_k$ ($k = 1, 2, \ldots\ldots N$) is 1, i.e., $\sum_{k=1}^{N} p_k = 1$. In the case of $p_k = 1$, there is no uncertainty for $x = x_k$. A lower value of p_k increases the uncertainty when it is known that $x = x_k$ occur, which also increases the information which is generally measured by $I(x_k) = -log(p_k)$. The information contained by the whole event set X is called entropy enumerated by the expected value of $-log(p_k)$, that is,

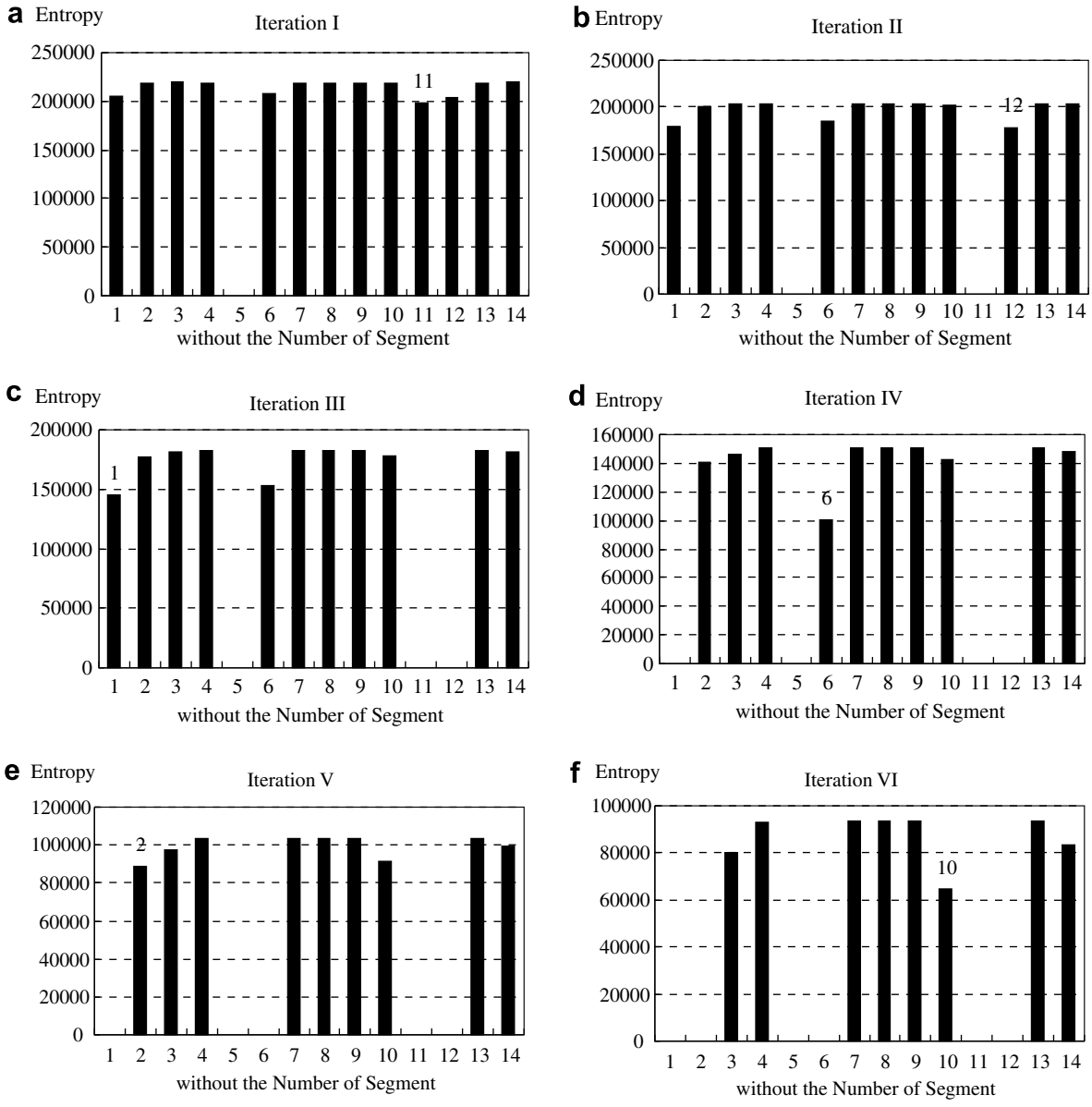

Fig. 4. Relationship of entropy when removing the number of segment for dataset "Property" iteratively.

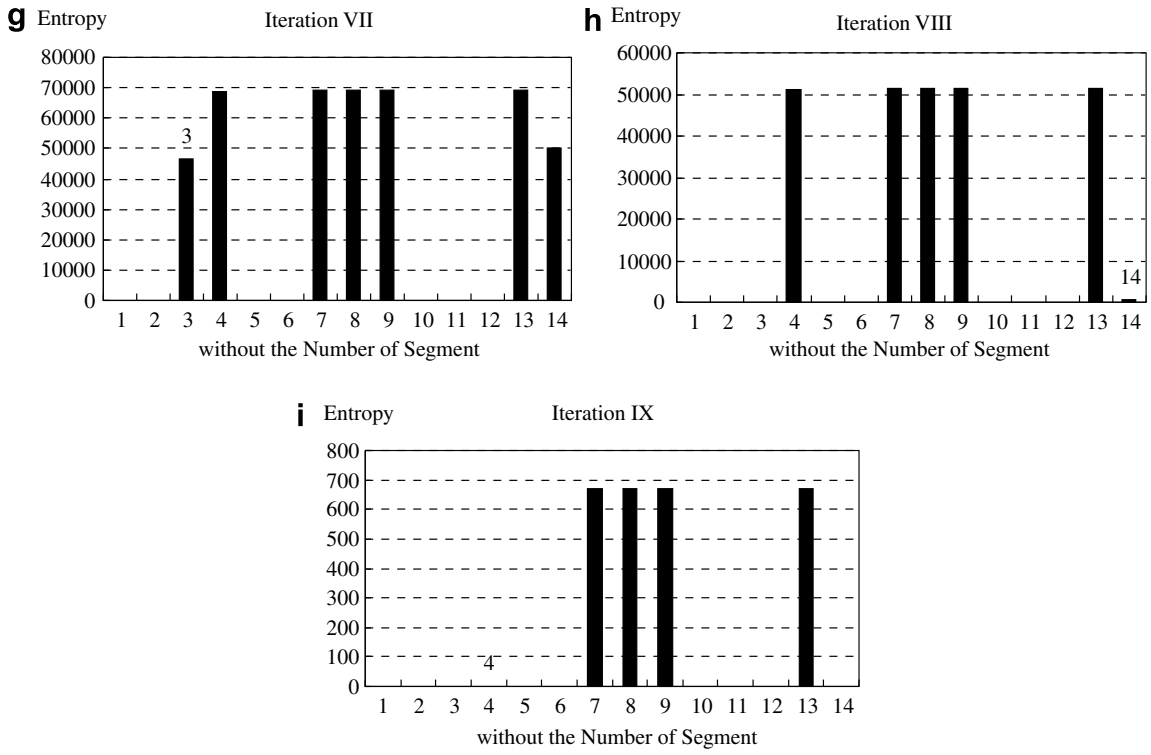

Fig. 4 (*continued*)

$$H(X) = E(I(x_k)) = -\sum_{i=1}^{N} p_i \log p_i$$

A large value of entropy $H(X)$ indicates a high uncertainty about X. When all the probabilities (i.e., p_k for all k) are equal to each other, the maximal uncertainty is obtained. The entropy $H(X)$ achieves its maximum $\log(1/N)$. On the contrary, when all the p_i except one are 0, there is no uncertainty about X, i.e., $H(X) = 0$.

Our proposed method is based on the process that the COA code will not be least affected when an irrelevant segment is removed.

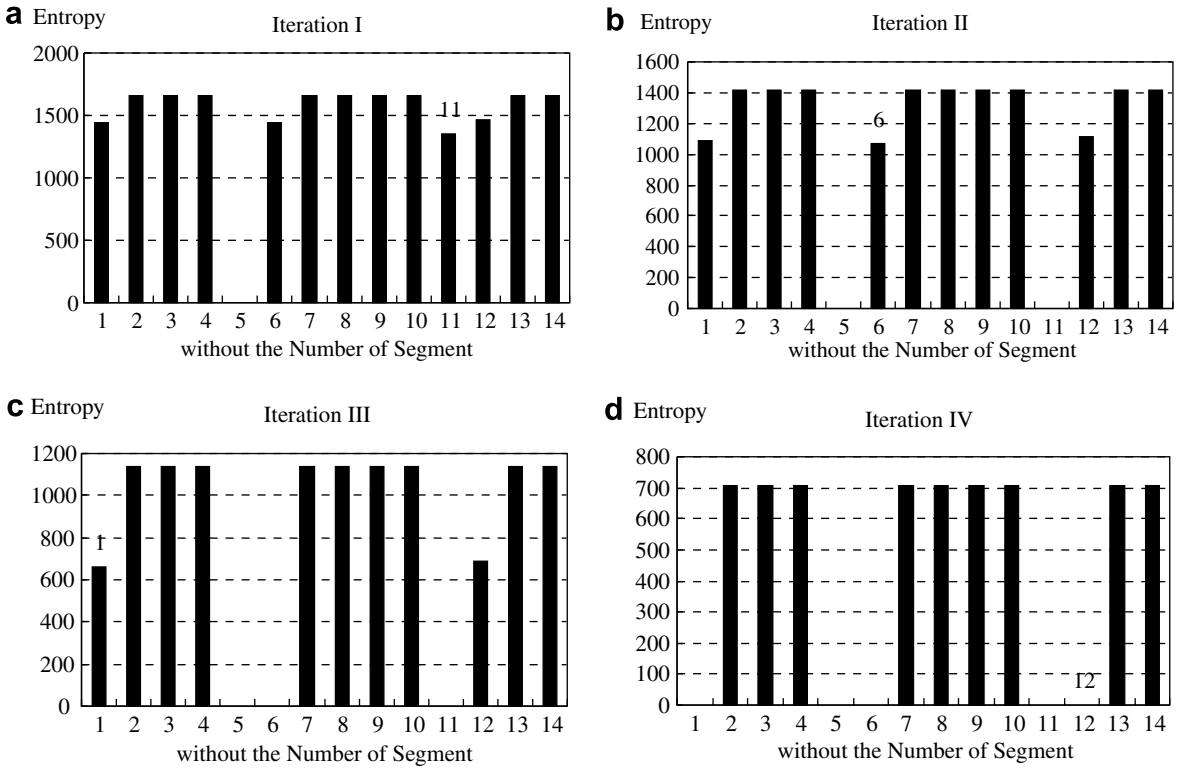

Fig. 5. Relationship of entropy when removing the number of segment for dataset "Stock" iteratively.

Removing one segment from the original code with nine dimensional space decreases the dimensionality of the whole dataset, which results in an eight dimensions. If the distinctness among these segments is close after projecting from a nine to eight dimensions, this segment is less relevant to the whole dataset. Applying SUD, features are ranked according to the entropy similarity measurement. Each feature in the feature subset is removed in turn to perform entropy calculation on the feature subset without that particular feature. The feature taken out from the feature subset that scores the minimum entropy is regarded as least important in the subset and is removed in the next iteration. Starting from the whole dataset with all the nine segments, the segment with the most distinctness among all the segments should be obviously removed first. According to the sequence, the segment with the second most distinctness is removed accordingly. However, from nine dimensions to eight dimensions, no matter which segment is removed, we get a new projection. There are altogether nine different projections and each projection has different segment distinctness. How to compare the projections is the core problem. By comparison, we can know which segment has the most distinctness.

We will say that the data has orderly configurations when it has distinct clusters while it has disorderly configurations when it has no distinct clusters (Dash et al., 1997). From the entropy theory (Fast, 1962) we have introduced above, if the configurations are orderly, the entropy should be low and if the configurations are disorderly, the entropy should be high. By the means of definition of entropy, we can measure that whether the segment is relevant or not. There are altogether eight segments for one COA code, which means that there are altogether nine iterations to decide the relevance of all the segments. In each iteration, the entropy is calculated under the condition of removing one of the segments. By comparing the entropy value of removing respective segment, the segment is removed as the most irrelevant segment since it gives the least entropy. This circulates until all the relevance of all segments is determined.

4. Simulation results

In this study, five real datasets obtained from an accounting firm are used to demonstrate the proposed method in establishing a COA structure. In order to simulate a more persuasive situation,

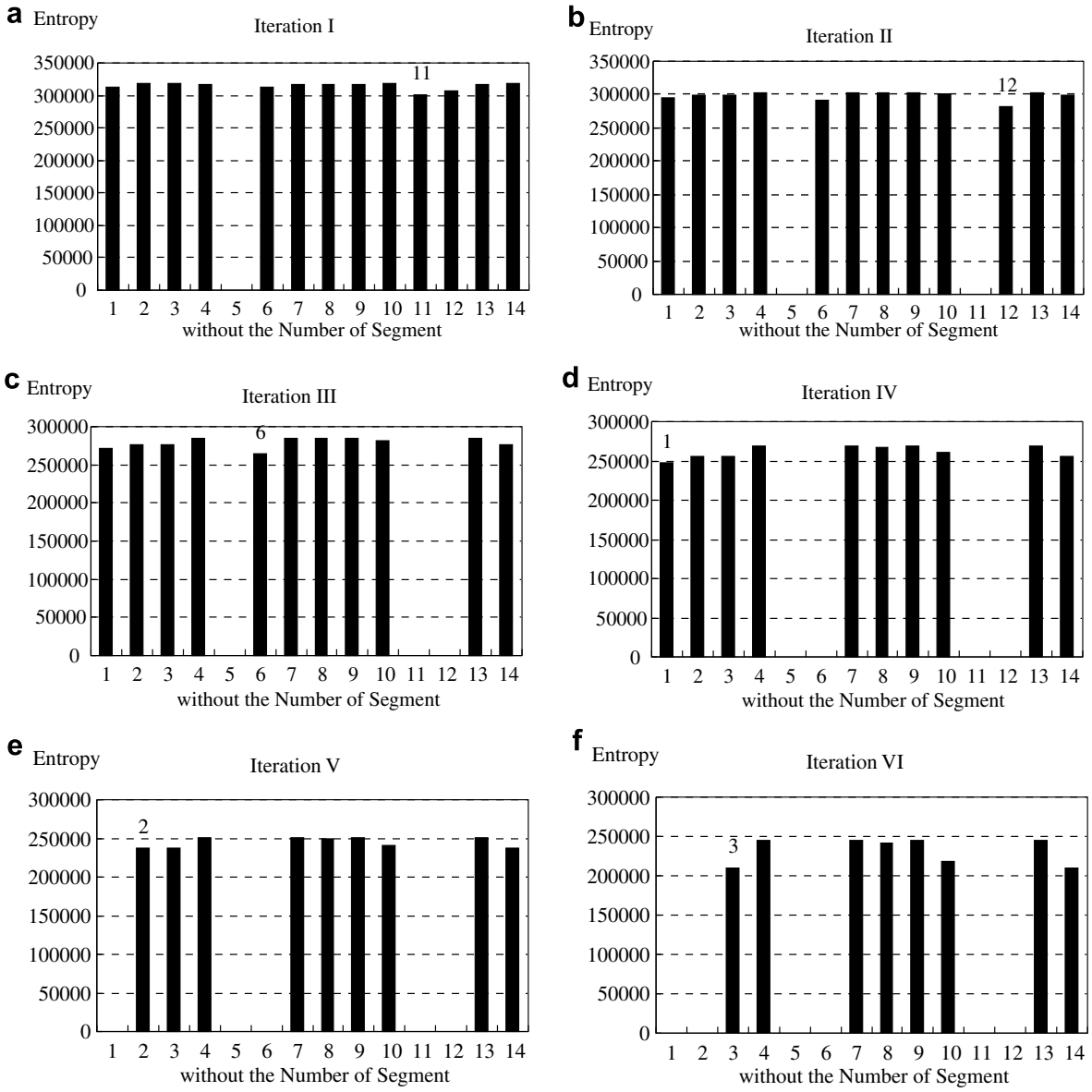

Fig. 6. Relationship of entropy when removing the number of segment for dataset "Video Film" iteratively.

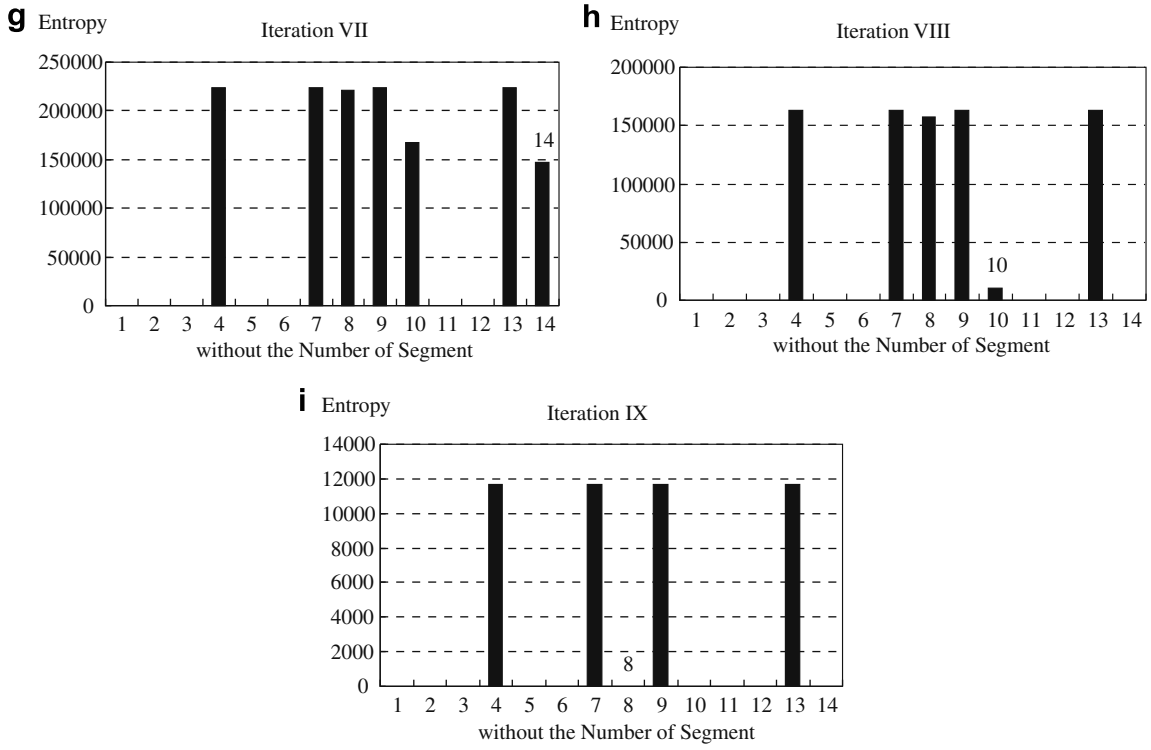

Fig. 6 (*continued*)

Table 4
Result of Entropy Measurement of Five Dataset

	1st removing feature	2nd removing feature	3rd removing feature	4th removing feature	5th removing feature	6th removing feature	7th removing feature	8th removing feature
Capital artist	Account section	Sub account	Activity	Cost center	Dept.	Intercompany	Section	Company
Investment	Account section	Company	Sub account	Activity	Dept.	Cost center	Section	Intercompany
Property	Account section	Sub account	Company	Activity	Dept.	Section	Cost center	Intercompany
Stock	Account section	Activity	Company	Sub account				
Video film	Account section	Sub account	Activity	Company	Dept.	Cost center	Intercompany	Section

class label of the five dataset is added in the fifth segment as the reference of different datasets. The class label is not included in the entropy measurement. The algorithm of determining the relevance of segments was written in Perl based on the platform of a desktop computer with 512M of RAM, Intel 43 GHz CPU Windows XP version 2002 Service Pack 2.

During the processing of the COA code, we have added five synthetic segments following the original eight segments. Although the five added segments are synthetic, they have different practical significance in the sense of accounting. The segments "Section", "Account Section", "Sub Account", "Business Nature", and "Intercompany" stand for the meanings from the 10th to the 14th segment. These five segments are not randomly generated data. In contrast, they are a part of used COA codes and have recently become obsolete after an overhaul of a COA structure. Thus, they can be considered as less relevant features compared with the segments which are currently still in used. We combine these thirteen segments together and regard them a new structure for entropy analysis. It is important to note that the physical nature of these segments will be used for analyzing the accuracy and reliability of the results.

Five datasets "Capital Artist", "Investment", "Property", "Stock" and "Video Film", are applied to the verification of the COA structure. Table 3 depicts the information of these five datasets. Dataset "Capital Artist" is discussed in the following paragraph and Fig. 2 gives the relationship between the respective entropy values when different segments are removed. The dataset "Capital Artist" consists of 14 features and 870 data instances with no missing data. They are real and reliable as they were all extracted from real accounting COA structures. With the nine features, which are the segments described in the COA structure, plus five synthetic features, the dataset has altogether fourteen features. We use SUD to determine which feature is to be removed. Fig. 2a–i illustrates that the segment with the lowest entropy is removed in each iteration. In Fig. 2a, it shows that the entropy values in the 1st iteration are more than 110,000 except the case when the 11th feature, "Account Section", is removed, because the entropy of the 11th feature is 108580.8. Thus, the 1st feature to be removed is "Account Section". After the feature "Account Section" is removed, thirteen features remain in the 2nd iteration. In the 2nd iteration as shown in Fig. 2b, the entropy is 98,101 when the 12th feature is removed, while the entropy is more than 104,000 when other features are removed. Thus, the 2nd most irrelevant feature is "Sub Account". It is worth noting that the first two irrelevant features are the synthetic features. This confirms the concept of the proposed methodology. From Fig. 2c–i, results from the 3rd iteration to the 9th iteration are presented. They show that Activity, Cost Center, Department, Intercompany, Section are the irrelevant features. It should be

noted that in Fig. 2i, the entropy is 0 after the feature "Account" is removed. Once when it reaches the level of zero-entropy, all entropy become zero in the following iterations making further calculation meaningless. In this study, the algorithm stops when it reaches zero entropy. Thus, apart from the synthetic features, 11, 12, 10, and 14, features "Activity", "Cost Center" and "Department" are removed from the COA structure. The remaining features then become the finalized COA structure.

Figs. 3–6 is the illustrations of the other four datasets, "Investment", "Property", "Stock" and "Video Film". Similarly, iteration by iteration is shown in each figure. There are 1058 data instances in dataset "Investment". From Fig. 3a, it can be found that the first feature to be removed is feature 11, "Account Section", because it is the only feature whose entropy is lower than 140,000. In the following iterations, "Account Section", "Company", "Sub Account", "Activity", "Department", "Cost Center", "Section" and "Intercompany" should be removed accordingly in the Fig. 3b–i. Table 4 presents the overall results inferred from Figs. 2–6. (Table 4) shows that features 10, 11, 12 and 14, which are all synthetic features, have been removed in the four datasets except "Stock". The difference between dataset "Stock" and the other four is mainly due to its relatively small data size. It is also noticed that the four real features, "Company", "Department", "Cost Center" and "Activity" have consistently been removed in the first several iterations for five datasets. These observations confirm our results.

The entropy distribution of all five datasets is analyzed in Table 4. In datasets, "Property", "Stock" and "Video Film", features "Company" and "Activity" have displayed similar values of entropy. This indicates one of them is a redundant feature that results in only feature "Company" being selected into the finalized COA structure. Similarly, the three datasets, "Capital Artist", "Investment" and "Video Film", show that either features "Department" or "Cost Center" is selected; feature "Department" is selected in our simulation. As shown in Table 2, a typical code of the finalized COA structure, which consists of seven segments, is read in an accounting sense of "*Company Sunny Ltd, IT department, sells equipment, to company Star Ltd, used by internal, in Hong Kong, Sunny Ltd. and Star Ltd are mutual-companies*". Compared to the original eight segments code that is read *"Company Sunny Ltd, IT department, sells hardware, of infrastructure, to company Star Ltd, used by internal, in Hong Kong, Sunny Ltd. and Star Ltd. are mutual-companies*", the above finalized code is sufficient to represent a transaction between companies.

Entropy quantifies the information contained in a message. In this application, entropy is used to determine the minimum seg-

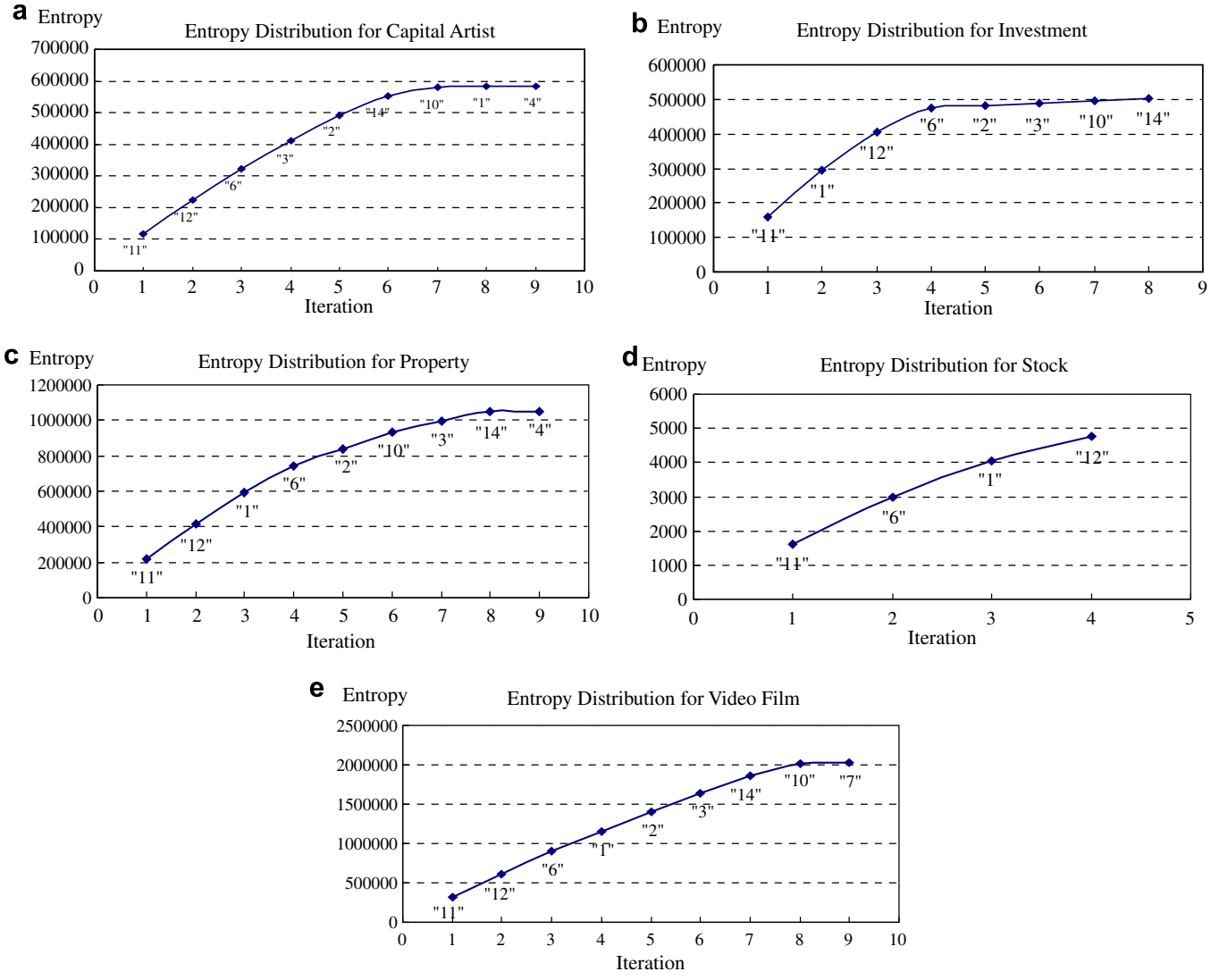

Fig. 7. Entropy curve for all the segments accumulated whenever removing one irrelevant segment for (a) capital artist, (b) investment, (c) property, (d) stock and (e) video film.

Table 5
Transformation of the Five Datasets in SOM Analysis

	Capital artist	Investment	Property	Stock	Video film
Instances	13	13	13	13	13
Dimensions	870	1058	1194	110	1416

ment length necessary to communicate information also represents a limit on the best possible lossless compression on the segments. Adding irrelevant segments to the COA structure does not have an effect on increasing the information as the content information is saturated. (Fig. 7) shows how entropy varies with the number of iteration for the five datasets. In Fig. 7a, dataset "Capital Artist", it shows the entropy increases after different segments are removed in different iterations. The entropy saturates at the 9th iteration, when segment "4" is removed. Like dataset "Capital Artist", Fig. 7b–e illustrate how entropy varies iteratively for the other four datasets.

Here, we analyze the COA structure from the clustering perspective. We describe how SOM, well-known with its visualization ability projecting high dimensional data into a two dimensional output SOM map, is used for analyzing the results. SOM is used because SOM output map is able to maintain the topological order of input data making SOM map a useful clustering method. In this study, all the data in the five datasets are projected into SOM maps.

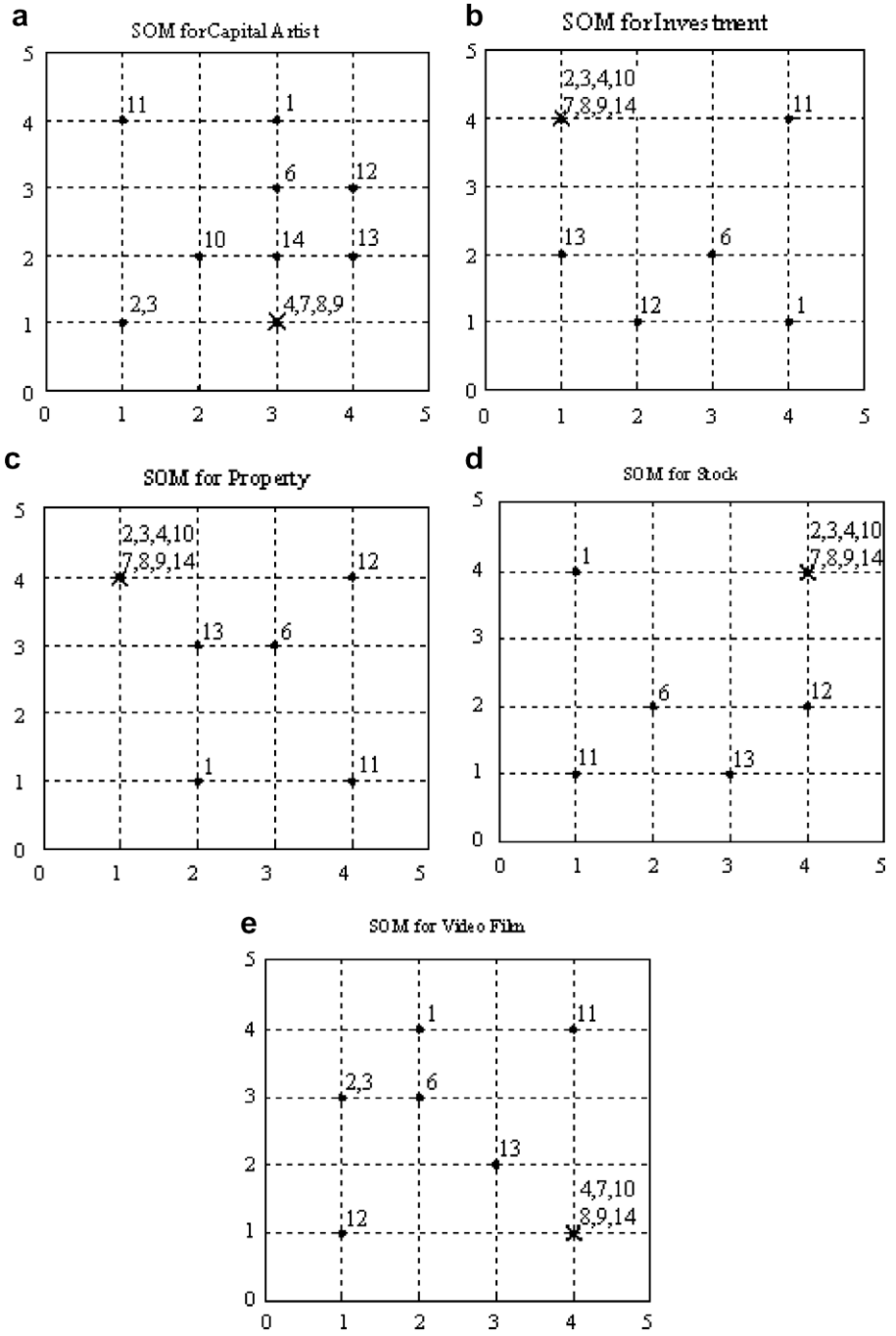

Fig. 8. SOM Projection of the segment clustering for respective dataset (a) capital artist, (b) investment, (c) property, (d) stock and (e) video film (The asteroid mean the biggest cluster in each SOM projection).

When projecting these datasets to SOM maps, the total number of instances in the given dataset is the dimension of the SOM input data, which results in thirteen data points with a high dimensional nature as shown in Table 5. In these five sets of data, removed segments or preserved segments are expected to exhibit certain extent of similarities that may be detected by the SOM. To find the relationship among the segments that are removed or preserved in the process of entropy calculation, the way how the SOM grouped the segments together provides very promising results. (Fig. 8) show the SOM output maps of the five datasets, namely "Capital Artist", "Investment", "Property", "Stock" and "Video Film". The clustering of Dataset "Capital Artist", for instance, is shown in Fig. 8a. Although there are a few segments showing sparse distribution, segment 2 and segment 3 are grouped together, and segments 4, 7, 8, 9 are grouped together as another cluster. This clustering result corroborates with the entropy result shown in Fig. 7a that segment 2 (Department) and segment 3 (Cost Center) are removed in the 5th and 4th iterations, respectively. And segments 4 (Account), 7 (Business Nature), 8 (Mutual Company) and 9 (Location) are selected as the finalized COA segments. In another example of "Video Film" dataset, Fig. 8e shows there are 2 clusters. The first one consists of segments 2 and 3, and the second cluster consists of segments 4, 7, 8, 9, 10 and 14, which corresponds to the entropy result displayed in Fig. 7e. Segments 2 (Department) and 3 (Cost Center) are removed in the 5th and 6th iterations. Segments 4 (Account), 8 (Mutual Company) and 9 (Location) are kept as the final COA segments, while segments 7 (Business Nature), 10 (Section), 14 (Intercompany) are more irrelevant compared with others. In Fig. 7b–d, they all have a cluster that consists of eight segments. In these cases, segments 2 (Department), 3 (Cost Center), 4 (Account), 7 (Business Nature), 8 (Mutual Company), 9 (Location), 10 (Section), and 14 (Intercompany) are found to be irrelevant from the corresponding results shown in Fig. 7. It is worth noting that the irrelevant segments are found in the same cluster that is different from the previous cases that selected segments are grouped in the same cluster. We must point out that SOM cannot determine the degree of relevance among the segments, because SOM only clusters segments, while SUD is able to determine the significance of each segment in terms of its relevance. The sparse clustering results as displayed in segments 1, 6, 11, 12 and 13 in Fig. 8 indicate these segments do not exhibit sufficient similarities. In summary, the SOM results corroborate the segment selection results based on SUD. Also, the finalized COA structures are found to be practically correct from an accounting perspective.

5. Conclusion

In this paper, a novel application of nominal data analysis to computational accounting system is introduced. "*Chart of Account*" (COA) is the structure of the Enterprise Requirement Planning (ERP) Accounting System. In order to maintain effective accounting operations, establishing a completely different COA structure is usually required in a space of few years. This is a costly and time consuming process. In this study, a COA structure is transformed as a large nominal dataset enabling nominal feature selection technique be used for determining the close-optimal number of accounting segments. Entropy measurement has been applied to COA using the method of SUD. The whole process is extremely cost effective and time efficient compared with the conventional accounting expertise demanding manual process. Our obtained results are promising from both computation and accounting perspectives. We also use Self-Organizing map to investigate the similarities between different segments. Although SOM, which only shows relevant or irrelevant segments clustering behavior, cannot be directly applicable for finding which segments to be selected, it is proved to be a useful approach in cross-examining the COA structure. At last, the study of this paper relies on real COA datasets, our obtained results corroborate that our computational results are correct from an accounting perspective.

References

Basak, J., De, R. K., & Pal, S. K. (1998). Unsupervised feature selection using neuro-fuzzy approach. *Pattern Recognition Letters, 19*, 997–1006.

Chow, T. W. S., & Huang, D. (2005). Estimating optimal features subset using efficient estimate of high dimensional mutual information. *IEEE Transactions on Neural Networks, 16*(1), 213–224 (January).

Chow, T. W. S., Rahman, M., & Wu, Sitao (2006). Content-based image retrieval by using tree-structured features and multi-layer self-organizing map. *Pattern Analysis and Application*.

Chow, T. W. S., & Wu, S. (2004). An online cellular probabilistic self-organizing map for static and dynamic data sets. *IEEE Transactions on Circuit and System I, 51*(4), 732–747.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Dash, M., Liu, H., & Yao, J. (1997). Dimensionality reduction of unsupervised data. In *Proceedings of ninth IEEE international conference on tools with artificial intelligence* (pp. 532–539).

Dash, M., Choi, K., Scheuermann, P., & Liu, H. (2002). Feature selection for clustering – a filter solution. In *Proceedings of second IEEE international conference on data mining* (pp. 115–122).

Dy, J. G., & Brodley, C. E. (2000). Feature subset selection and order identification for unsupervised learning. In *Proceedings of the seventeenth international conference on machine learning*.

Fast, Johan Diedrich (1962). *Entropy: The significance of the concept of entropy and its applications in science and technology. The statistical significance of the entropy concept*. Eindhoven: Philips Technical Library.

Huang, D., & Chow, T. W. S. (2004). Effective feature selection scheme using mutual information. *Neurocomputing, 63*, 325–343 (August).

Huang, D., & Chow, T. W. S. (2005). Efficiently searching the important input variables using Bayesian discriminant. *IEEE Transactions on Circuit and Systems, Part I, 52*(4), 785–793 (April).

Irv Chasen. How to start and maintain a cost accounting system. *Painting & Wallcovering Contractor* (Vol. 68, p. 34). St. Louis: Mar/Apr 2006.

Kohonen, T. (1995) (2nd ed.). *Self-organizing maps. Series in information sciences* (Vol. 30). Heidelberg: Springer. 1997.

Magnusson, C., Arppe, A., Eklund, T., Kloptchenko, A., Back, B., Visa, A., et al. (2005). Combining collocational networks and self-organizing maps in analyzing quarterly reports. *Information and Management, 42*(4), 561–574.

Martín-del-Bŕıo, B., & Serrano-Cinca, C. (1993). Self-organizing neural networks for the analysis and representation of data: Some financial cases. *Neural Computing and Applications, 1*(3), 193–206.

Mitra, P., Murthy, C. A., & Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(3), 301–312.

Oracle General Ledge User Guide (2003). Release 11i (pp. 5-7).

Enterprise Resource Planning, Deciding the Accounts Chart Mukul Pareek (2000).

Shi, S. Y. M., & Suganthan, P. N. (2003). Unsupervised similarity-based feature selection using heuristic Hopfield neural networks. In *Proceedings of the international joint conference on neural networks* (Vol. 3, pp. 20–24).

Wu, S., & Chow, T. W. S. (2005). PRSOM – a probabilistic regularized self-organizing map for data projection and visualization. *IEEE Transactions on Neural Networks, 16*(6), 1362–1380.

Yang, Y., Guan, X., & You, J., 2002. CLOPE: A fast and effective clustering algorithm for transactional data. In *Proceeding of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 682–687).

Yun, C. H., Chuang, K. T., & Chen, M. S. (2001). An efficient clustering algorithm for market basket data based on small large ratios. In *Proceedings of twenty-fifth annual international computer software and applications conference* (pp. 505–510).