Journal of Southeast University (English Edition) Vol. 24, No. 3, pp. 354 – 357 Sept. 2008 ISSN 1003—7985

Clustering analysis algorithm for security supervising data based on semantic description in coal mines

Meng Fanrong Zhou Yong Xia Shixiong

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: In order to mine production and security information from security supervising data and to ensure security and safety involved in production and decision-making, a clustering analysis algorithm for security supervising data based on a semantic description in coal mines is studied. First, the semantic and numerical-based hybrid description method of security supervising data in coal mines is described. Secondly, the similarity measurement method of semantic and numerical data are separately given and a weight-based hybrid similarity measurement method for the security supervising data based on a semantic description in coal mines is presented. Thirdly, taking the hybrid similarity measurement method as the distance criteria and using a grid methodology for reference, an improved CURE clustering algorithm based on the grid is presented. Finally, the simulation results of a security supervising data set in coal mines validate the efficiency of the algorithm.

Key words: semantic description; clustering analysis algorithm; similarity measurement

With the fast development in information-based constructions, coal mines produce a large number of data resources appearing in the forms of text, drawings, numerical data, audio data and images, which contain a richness of information resources related to production and life. Ref. [1] presented a semantic description model of complex multi-source data in coal mines and the hybrid similarity measurement, it solved the elementary problem of how to describe complex multi-source data in coal mines. But how to discover the useful information about safety production and management which can guide production safety and enhance economic efficiency from the obtained security supervising data in coal mines becomes very necessary.

In view of the above, combined with the clustering analysis method, this paper presents a clustering analysis algorithm for security supervising data based on semantic description in coal mines. The corresponding description of the algorithm is given and simulation results of the algorithm validate its efficiency.

1 Semantic and Numerical-Based Description of Security Supervising Data in Coal Mines

Ref. [1] presented a description model of security supervising data based on semantic and numerical data in coal mines, which not only overcame the boring character of the description by numerical data alone, but also avoided the complex character of the description only by semantics. So it presented the coal mine data's character of multi-source and complex and provided a powerful tool for further in-depth study. The description model used a tree structure to describe and obtain the location of numerical concepts and their relationships. Meanwhile, the nodes in the tree were attached with some numerals which denoted different attribute values at this location. The specific structure is shown in Fig. 1[1].

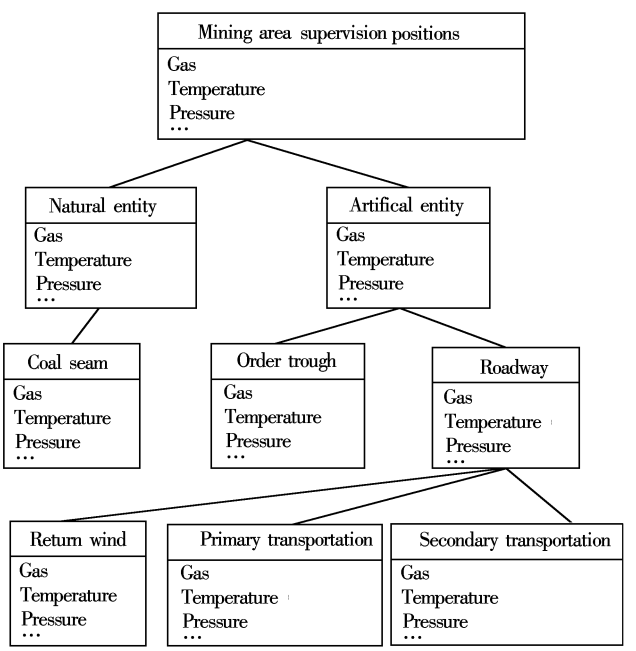

Fig. 1 Semantic-and numerical-value-based data description

2 Semantic and Numerical-Based Hybrid Similarity Measure

At present, the semantic similarity research is divided into two groups. One refers to the semantic-distance-based measure, where the shorter the semantic distance between two concepts within the dendriform structure is, the more similar the concepts are; while the other is based on the information quantity of the two concepts, where the more shared information they have, the more similar they are[2].

Ref. [1] used the second way to measure the semantic similarity in the semantic description of the supervising position in coal mines. As Resnik et al. [3] pointed out, the more shared information two concepts have, the more semantically similar they are; this can be described as follows[1-3]:

$$\mathrm{sim}(c_i, c_j) = \max_{c \in S(c_i, c_j)} [-\log P(c)] \quad (1)$$

Received 2008-04-15.

Biography: Meng Fanrong (1962—), female, doctor, associate professor, mengfr62@163.com.

Foundation items: The National Natural Science Foundation of China (No. 50674086), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20060290508), the Postdoctoral Scientific Program of Jiangsu Province (No. 0701045B).

Citation: Meng Fanrong, Zhou Yong, Xia Shixiong. Clustering analysis algorithm for security supervising data based on semantic description in coal mines[J]. Journal of Southeast University (English Edition), 2008, 24(3): 354 – 357.

where $P(c)$ = count/sum, count represents the number of subsets of concept c and sum refers to the number of the concepts belonging to the same ontology. Thus, $-\log P(c)$ expresses concept c's information quantity and $S(c_i, c_j)$ stands for the collection of concepts of the original positions of c_i and c_j[3].

However, Eq. (1) presents the semantic similarity ignoring the concept information quantity, so its characterization is not comprehensive enough. Based on Eq. (1), Lin put forward a new similarity measurement which has a relationship with not only their common origin but also with the quantity of information they contain, described as[4]

$$\mathrm{sim}(c_i, c_j) = \frac{2 \times \max\limits_{c \in S(c_i, c_j)} [-\log P(c)]}{\log(P(c_i) + P(c_j))} \tag{2}$$

where $P(c)$, $P(c_i)$ and $P(c_j)$ refer to the occupancy of the subsets within the same ontology, respectively.

For the similarity measure of the numerical data, the general Minkowski distance is selected among various measures to describe the numerical similarity of the multi-dimensionality in coal mine data. It can be denoted as follows[1]:

$$L_k(X, Y) = \left(\sum_{i=1}^{d} |x_i - y_i|^k \right)^{1/k} \tag{3}$$

where $X = (x_1, x_2, \ldots, x_d)$ and $Y = (y_1, y_2, \ldots, y_d)$ refer to the d-dimensional data in coal mines to be supervised respectively while k is a positive integer.

Through the above conceptions of semantic similarity measures and numerical similarity measures, it is not difficult to obtain an effective measurement which combines these two similarity measures. The Minkowski distance, such as organic complementarities to numerical value measures, is applied in computing dimensional properties similarity of data in coal mines. It can be of help to describe the similarities among different position concepts via the semantic-based similarity formula as

$$d(R_x, R_y) = \alpha L_k(X, Y) + (1 - \alpha)\,\mathrm{sim}(c_x, c_y) \tag{4}$$

where $L_k(X, Y)$ and $\mathrm{sim}(c_x, c_y)$ are computed respectively via Eq. (1) and Eq. (3), $0 < \alpha < 1$, accounting for the weight occupancy of the numerical value property and the position concepts within coal mine data similarity[1].

3 Clustering Algorithm for Security Supervising Data Based on Semantic Description in Coal Mines

Based on the similarity measure of semantic description of the security supervising data which has been discussed in section 2, such as a basic distance criteria for clustering, this paper makes use of the improved CURE algorithm for clustering analysis for semantic description of the security supervising data. In order to improve efficiency, with the characteristics of the security supervising data in coal mines, it can use the gird algorithm as a reference. First, data space is divided into some small regions, and each region is treated as an initial clustering whose number of objects in it has reached a certain amount, but the rest of data is still the same as those of the separate categories. After this regional division, the clustering is accomplished by implementing other steps of the CURE algorithm[5].

3.1 Division of the original categories

D is a d-dimensional data space, and each dimension of D is divided into $k_1, k_2, \ldots, k_d$ regions; namely, the i-th dimension is divided into $k_i (1 \leqslant i \leqslant d)$ parts. Then a regional cell can be denoted by a d-dimensional array as cell = $C[s_1][s_2]\ldots[s_d]$, where s_i is the s_i-th region in the i-th dimension, and the value of s_i is between 0 and $k_i - 1$. We consider that $\mathrm{cell}_1 = C_1[i_1][i_2]\ldots[i_d]$ and $\mathrm{cell}_2 = C_2[j_1][j_2]\ldots[j_d]$ are neighbors if they can satisfy the following conditions[5]:

$$\begin{cases} |i_p - j_p| = 1 & p = v \ (1 \leqslant v \leqslant d) \\ i_p = j_p & p = 1, 2, \ldots, v, v+1, \ldots, d \end{cases}$$

According to the two-dimensional data space and the three-dimensional data space, it is considered cursorily that there are 2-D adjacent cells in the d-dimensional data space. However, not only the number of regions is greatly reduced, but also the increasing relationship among the number of regions and dimensions are changed from an exponential growth into a linear growth, so it can reduce the time required for running the algorithm.

3.2 Algorithm description

3.2.1 Dividing data space

Scanning the data set, the dividing intervals of each dimension according to the average distance of the adjacent data points are obtained. The processes for dividing each dimension are as follows:

1) Sort the data points, and obtain the ranges for each dimension.

2) Obtain the intervals between adjacent data points, and record each interval value and the number of times they occur.

3) Obtain the dividing interval, which is an integer nearest to the value calculated by the following formula:

$$W_{s_i} = \mathrm{round}\left(\frac{\sum\limits_{t=1}^{l_i} \ln v_t m_t}{\sum\limits_{t=1}^{l_i} m_t} + b_i z \right)$$

where s_i is the dimensional ID, W_{s_i} is the dimensional dividing interval; l_i is the number of different dividing intervals in s_i. The value of each interval and the corresponding number is in v_t and m_t, where b_i is the step value for the i-th dimension and z is a manual parameter.

3.2.2 Establishing Hash table H

Scan the dataset for a second time and record the regional information where each data point is in the Hash table H. When each data point is recorded, the regional cell of this point should be calculated first, and then the corresponding regional cell and the data point are mapped into the table H using the Hash function. If there is no record in a regional cell of this data point, the relevant information will be put into table H. The corresponding count 1 and count 2 will be counted as well. If the regional cell of this data point has been recorded in table H, we will just make an operation on the corresponding count 2.

3.2.3 Judging whether regional cell is empty

Check the count of every regional cell in the Hash table; judge whether it is an empty regional cell, and mark the empty regional cell. After the judgment of all the regional cells, all the empty regional cells can be deleted.

3.2.4 Data structure conversion

After completing the partition of a region, we need to add the information of the initial clusters stored in the regional cell and the single cluster appearing as an individual into the k-d tree and the heap Q of CURE.

3.2.5 Continue to implement the CURE algorithm

On the basis of these initial clusters and the remaining single clusters, the CURE algorithm continues to be implemented.

4 Simulations

In order to analyze the performance of the improved algorithm conveniently, this paper selects security supervising data in coal mines as experimental objects which are listed in Tab. 1. One group of data is of two dimensions including gas chroma and the location of gas sensors. The other group is of four dimensions which still contains temperature and the position of temperature sensors as well as gas chroma and the location of gas sensors[6]. The results comparing the data mining performance of CURE with the improved CURE are given in Tab. 2 and Tab. 3.

Tab. 1 The status of dataset

Data dimension	Records	Attribute
Two dimension	38 805	Gas chroma and the location of gas sensors
Four dimension	53 255	Gas chroma and the location of gas sensors, temperature and the position of temperature sensors

Tab. 2 The results of two-dimensional data ($k = 5$)

Algorithm	Cluster			
	Cluster one	Cluster two	Cluster three	Outlier
CURE	9 773(25.18%)	15 213(39.20%)	13 727(35.37%)	92(0.24%)
Improved CURE	11 131(28.68%)	14 519(37.42%)	13 107(33.78%)	48(0.13%)

Tab. 3 The results of four-dimensional data ($k = 5$)

Algorithm	Cluster			
	Cluster one	Cluster two	Cluster three	Outlier
CURE	12 350(23.19%)	36 213(68%)	4 521(8.49%)	171(0.32%)
Improved CURE	9 549(17.93%)	39 289(73.78%)	4 327(8.13%)	90(0.17%)

Comparing the results of the two algorithms, we can obtain that the three major clusters are relatively stable when $k = 5$, while the other two small clusters can be considered as some noise data which is not eliminated completely because of their small proportion. Thereby, the clustering results will be able to reflect the real situation when $k = 5$. Three major clusters represent three kinds of distribution of supervising data, while the other two small clusters can be ignored. It does not affect the final analysis. Combined with the relevant experience and knowledge of coal mine fields, three clusters can be defined as high-risk datasets (28.68%), normal datasets (37.42%), and low-risk datasets (33.78%), respectively, after inspecting the bounds of the three kinds of data. High-risk datasets indicate that in working face the following phenomena, such as poor underground ventilation, the gathering of corner gas, abnormal gas prominence, increased local temperature and so on, possibly occur. Under such circumstances, it can lead to the occurrence of major incidents if there is mechanical or electrical fire, blasting open flame or other sources of fire. However, the normal datasets and low-risk datasets have little opportunity to occur in such situations. Consequently, we should pay particular attention to the high-risk datasets and take all necessary measures to reduce the emergence of such datasets.

5 Conclusion

In allusion to the characteristics of the security supervising data based on semantic description in coal mines, first, the semantic and numerical-based hybrid description method of security supervising data in coal mines is described. Secondly, the similarity measurement method of semantic and numerical data are given and a weight-based hybrid similarity measurement method for the security supervising data based on the semantic description in coal mines is presented. Thirdly, taking the hybrid similarity measurement method as the distance criteria and using a grid methodology for reference, an improved CURE clustering algorithm based on the grid is given. Finally, the simulation results of a security supervising data set in coal mines validate the efficiency of the algorithm.

References

[1] Zhou Yong, Xia Shixiong. A hybrid similarity measurement for complicated and multi-source data in coal mine [J]. *Journal of Jiangnan University*: *Natural Science Edition*, 2007, **6** (6): 665 – 668. (in Chinese)

[2] Alexander M, Steffen S. Measuring similarity between ontologies [C]//*Proceedings of the* 13*th International Conference on Knowledge Engineering and Knowledge Management*. Springer-Verlag, 2003: 251 – 263.

[3] Resnik P. Using information content to evaluate semantic similarity in taxonomy [C]//*Proceedings of the International Joint Conference on Artificial* 14*th Intelligence*. Montreal, Canada, 1995: 445 – 453.

[4] Lin D. An information-theoretic definition of similarity [C]//*Fifteenth International Conference on Machine Learning* (*ICML*'98). Madison, Wisconsin, 1998: 296 – 304.

[5] Cao Hongqi, Yu Lan, Sun Zhihui. An algorithm of outliers mining based on grid clustering techniques[J]. *Computer Engineering*, 2006, **32**(11): 119 - 121. (in Chinese)
[6] Yang Yanguo, Ti Zhengyi. Research on fuzzy comprehensive evaluation method of mine coal and gas outburst hazard [J]. *Express Information of Mining Industry*, 2006, **32**(8): 33 - 36. (in Chinese)

基于语义描述的煤矿安全监测数据聚类分析算法

孟凡荣　周　勇　夏士雄

(中国矿业大学计算机科学与技术学院,徐州 221116)

摘要:为了挖掘基于语义描述的煤矿安全监测数据中蕴含的生产安全信息,指导煤矿安全生产和决策,研究了基于语义描述的煤矿安全监测数据聚类分析算法. 首先,阐述了煤矿安全监测数据的语义和数值混合描述方法;接着,分别给出了语义和数值数据的相似性度量方法,以及基于权重的煤矿安全监测数据的混合相似性度量方法;然后,以混合相似性度量方法为距离度量准则,并借鉴网格的思想,给出了基于网格的改进 CURE 聚类算法. 通过煤矿安全监测数据集的仿真实验,验证了所提算法的有效性.

关键词:语义描述;聚类分析算法;相似性度量

中图分类号:TP18