Visual Data Mining Using a Constellation Graph

Tokihiko Niwa*, Kenji Fujikawa**,
Kazuyoshi Tanaka**, and Mayumi Oyama***
* Kwansei Gakuin High School
Uegahara 1-1-155, Nishinomiya, 662-8501 Japan
**Web System Support Center, Hitachi Systems and Services, Ltd.
***Kwansei Gakuin University, Center for Information & Media Studies
Uegahara 1-1-155, Nishinomiya, 662-8501 Japan
niwa@kwansei.ac.jp, ke-fujikawa@hitachi-system.co.jp,
kzy-tanaka@hitachi-system.co.jp, oyama@kwansei.ac.jp

Abstract. Software was developed to enable visual data mining of multivariate data using a constellation graph. Two applications of the new software are presented. A constellation graph is an effective way to display multivariate data on to a two dimensional plane. The value of the classification variable used just as it is or it classifies it into some levels. Then, the cluster on each level is made to be able to be identified using the color or the symbol. The operator can change the weight of the variable interactively and do the visual mining by seeing the change of the star on the graph. Moreover, our program computes the weight of the variable to separate each cluster. It is possible to presume the data, which cluster it belongs to. In addition, using a computer mouse, the operator can select an area of the graph to examine data in detail.

1 Introduction

The aim of this research is shown below:

1. It displays multivariate data using the constellation graph on the 2-dimension plane.

2. By changing weighting factor to each variable, being interactive from 0 to 1, the position and the movement of the star on the constellation graph can be examined.

3. Choosing one classification variable and classifying it into the level to distinguish between each level using the color and the symbol on the graph.

4. The condition and the change of the cluster, which has a value with the same level on the constellation graph, can be examined by changing the weight.

5. Calculating the weighting factor so that the within-cluster variance has the minimum value and inter-cluster has the maximum value.

6. Cutting off the cluster and the necessary part from the graph and it preserves them in the file and they can be used for another analysis.

7. By displaying the data that a classification variable is not proved in the same sample data on the constellation graph, the level can be presumed.

Visual data mining using a constellation graph isn't suitable for most data processing applications, due to the limitations of the graphical display. However, since a constellation graph is an effective way to display multivariate data onto a two dimensional plane. The program that we developed allows the operator to interactively change the weighting factors of the variable. As these values are changed, the display is automatically updated so that the operator can immediately see how changes affect the clustering of the data.

The weighting factor that is given to each variable determines the length of the vector for each variable drawn on the constellation graph. If it is not possible to position a star in a desired position, the operator must reconsider the data in question to determine whether there is a more effective way to mine the data. When a cluster of stars becomes very cluttered, it is possible to select that area for further investigation using a computer mouse. The data from this area is then saved in a new file that can be investigated separately. This has proven to be a very effective technique for visual data mining. Another methods to display multivariate data have been discussed [1] [2]. However, there were problems when the number of the variable increased, and it was not possible

to change the weighting factors. Especially, in case of the parallel graph, the pattern is different when changing the order of the display of the variable but the constellation graph isn't related with the order of the variable because it is a vector. And the number of the variable can increase the capacity of the computer as far as it permits.

2 The Constellation Graph

This chapter shows how to construct the constellation graph and how to calculate the optimum weighting factors so that the within-cluster variance has the minimum value and the inter-cluster variance has the maximum value on the constellation graph.

3.1 2.1 Constructing a Constellation Graph

A constellation graph uses multivariate table-type data. The output of this data is the observed value of a star on the half pie chart of the constellation graph. The stellar position is determined by an angle that is dependent on the variable and by a vector that is dependent on the weight of each variable. A connected graph is made and a star is displayed at the end of the graph. This is called a "constellation graph" as it is made up several stars, each representing a portion of the data. Constellation graphs were originally proposed in 1977 [3]. The means by which a general constellation graph is constructed is described in the Appendix.

3.2 Constructing a Constellation Graph with Weighted Variable

Having chosen a classification variable, the values of the weighting factors to be used are then based on either "what to read in the data" or "how to read the data". The cluster means the category value or the repartition value of the classification variable. To sort the data one can:

1. Sort by gathering clusters of related data, or
2. Sort using regression curves that were made by the clusters

Our program uses the first method. To accomplish this, we must first choose an average mark as a basing point and then determine the value of the average mark.

The classification variable n is divided into kinds of clusters. Next, we consider the terminal coordinates of the connection vector of the variable that consist of j kinds of variables, here equations (14) and (15) on Appendix are used.

$$\left(\xi_{ij}, r_{ij}\right)\left(i = 1,2,\cdots, m_j \quad j = 1,2,\cdots, n \quad N = \sum_{k=1}^{n} m_k \right) \qquad (\mathbf{1})$$

This distinguishes similar clusters of data, while at the same time ensuring that there is very little overlap between clusters. Equations (2) and (3) are used to determine the average of each cluster.

$$\overline{C_j} = \frac{1}{m_j} \sum_{i=1+m_{j-1}}^{m_j} r_{ij} \cos\xi_{ij} \quad , \quad \overline{S_j} = \frac{1}{m_j} \sum_{i=1+m_{j-1}}^{m_j} r_{ij} \sin\xi_{ij} \qquad (j = 1,2,\cdots, n \quad m_0 = 0) \qquad (\mathbf{2})$$

$$\overline{\xi_j} = \tan^{-1} \frac{\overline{S_j}}{\overline{C_j}} \qquad (j = 1,2,\cdots, n), \qquad \overline{R_j} = \sqrt{\overline{C_j^2} + \overline{S_j^2}} \qquad (j = 1,2,\cdots, n) \qquad (\mathbf{3})$$

The whole average mark is determined using equations (4) and (5).

$$\overline{C} = \frac{1}{N} \sum_{j=1}^{n} \sum_{i=1+m_{j-1}}^{m_j} r_{ij} \cos\xi_{ij}, \qquad \overline{S} = \frac{1}{N} \sum_{j=1}^{n} \sum_{i=1+m_{j-1}}^{m_j} r_{ij} \sin\xi_{ij} \qquad (j = 1,2,\cdots, n \quad m_0 = 0) \qquad (\mathbf{4})$$

$$\overline{\xi} = \tan^{-1} \frac{\overline{S}}{\overline{C}} \qquad (j = 1,2,\cdots, n), \qquad \overline{R} = \sqrt{\overline{C}^2 + \overline{S}^2} \qquad (j = 1,2,\cdots, n) \qquad (\mathbf{5})$$

Merdia [4] and Fisher [5] defined this average using circle coordinates. Our program uses this for the value of the average mark. Based on this average mark, it then calculates each data point and a vector to the average mark at the top of the circumference in the shooting shadow. The radius is from the center of the segment that links the starting point and an average mark with this circumference. Then, it sums the data in every cluster and the distance at which there was a shooting shadow using the average mark. In addition, it sums the distance between all the data and the average mark using equation (6).

$$Var_j = \sum_{i=1+m_{j-1}}^{m_j} \overline{R_j}(\xi_i - \overline{\xi_j}) \quad (j = 1,2,\cdots,n) \qquad Var = \sum_{j=1}^{n} \sum_{i=1+m_{j-1}}^{m_j} \overline{R}(\xi_i - \overline{\xi}) \qquad \textbf{(6)}$$

with $m_0 = 0$.

Now, we want to scatter the big clusters rather than the small ones. In other words, we must decrease the value of Var_j and increase the value of Var. To do this, we define a value of J that is determined by equation (7). To determine the best weighting factors to cluster every cluster, J should be minimized.

$$J = \sum_{j=1}^{n} \frac{Var_j}{m_j} \cdot \frac{N}{Var} \qquad \textbf{(7)}$$

3 Visual Data Mining using a Constellation Graph

The following example illustrates the visual data mining technique using a constellation graph. The example data that we have chosen are the familiar Iris data, which include five variables (the kind of iris, petal length, petal width, sepal length, and sepal width). There are three kinds of iris. The kind of iris is used as the dependent value and is colored with three different colors on the graph. The other values are used as the variables.

The constellation graph visual data mining system has two display forms. One is a data table view and the other is a graph view.

3.3 The Data Table View

The data table view is shown Figure 1. The input data can be loaded into the table from the keyboard or from a file saved in CSV format.

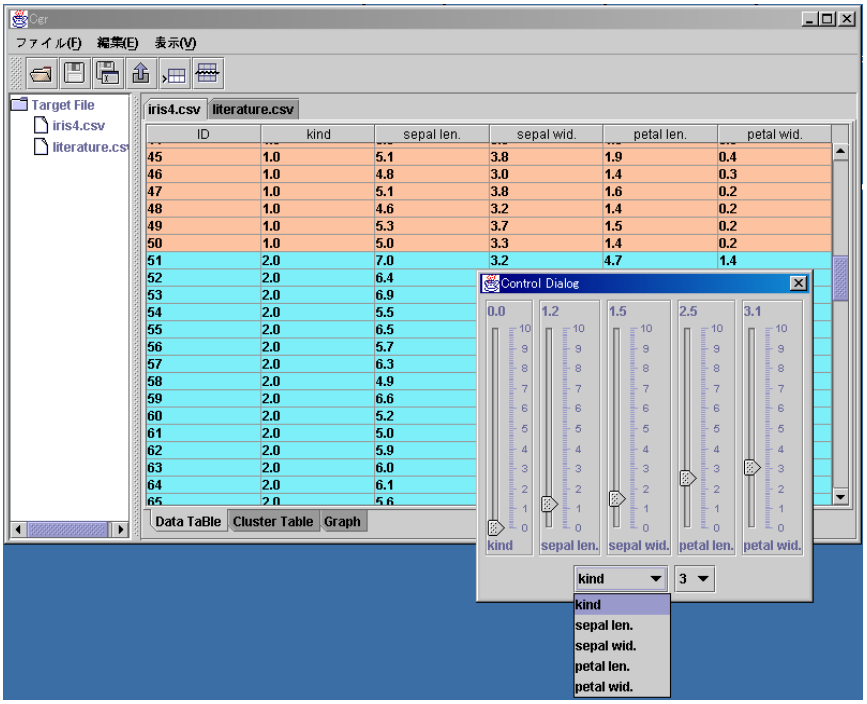

Fig. 1. Each line in the table is a variable. The variable at the left end of the line assigns the ID number of the data, which shows the number of the observation. Each row consists of the values of each observation.

Once the data input has been completed, the control dialog that displays the weight of each variable is displayed. Using the slide bars on the control dialog display and the mouse, the operator can change the weight of each variable. One variable is selected as the classification variable. Once the classification variable is chosen, the data table is sorted based on the range of the other variables. Observations with different values of the classification variable are displayed in different colors. For example, in Figure 1, the classification variable is “kind”, and three different colors are used to show the three

different kinds of iris. The operator can change the choice of the classification variable at any time. In the graph view, the color of the star changes to show the different values of the classification variable.

3.4 The Graph View

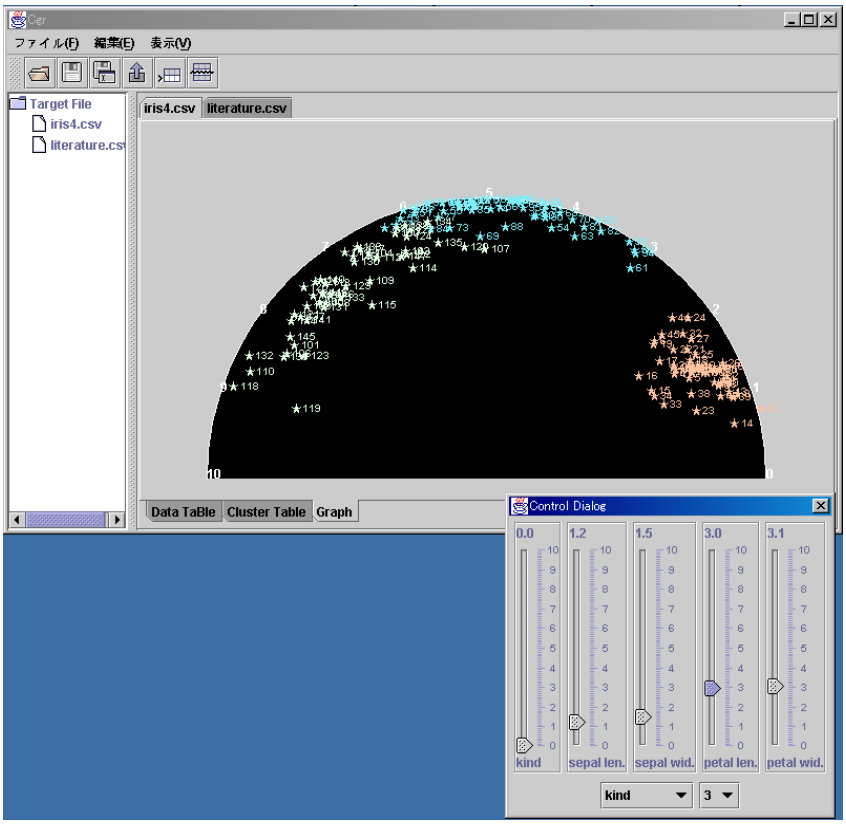

Fig. 2. The graph view (petal length=1.2, petal width=1.5, sepal length=3.0, sepal width=3.1)

The constellation graph is shown in Figure 2. The control dialog display can be used to change the values of the variables in the same way as used for the data table display. The graph display updates automatically in real-time as the value of a variable is changed. The different values of the classification variable are shown as different colors on the graph. Since the weighting factor of the classification variable is fixed at zero, it is not used in calculating the stellar position on the graph. Figure 3 shows the graph view changing the weighting factor of variable.

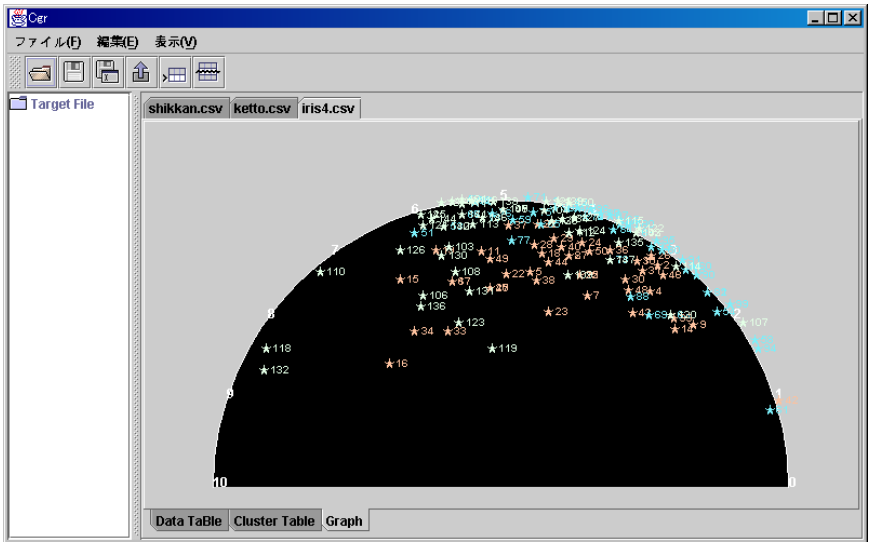

Fig. 3. The graph view (petal length=1.2, petal width=1.5, sepal length=0.0, sepal width=0.0)

3.5 Extracting Data from the Constellation Graph

As shown in Figure 4, using the graph view, the operator can use the mouse to choose a range of data designated by a polygon. This selected range can be deleted or saved in a separate file for later analysis. Data extracted and saved from a graph are stored in CSV format. Therefore, the data file can be easily used as the input for other statistical analysis and rule discovery tools.

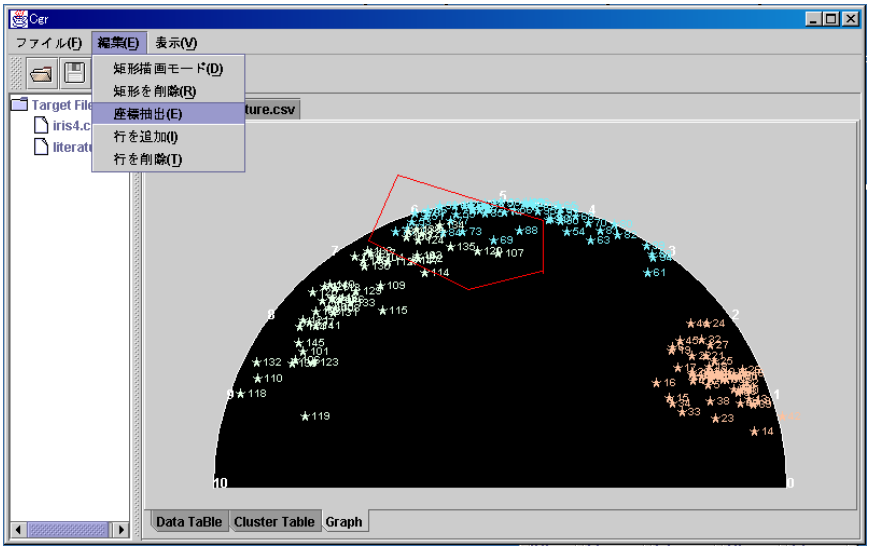

Fig. 4. Extracting data from the constellation.

4 Visual Data Mining using a Constellation Graph

To illustrate data mining using a constellation graph, two examples are discussed in the following sections.

3.6 An Example that Identified the Characteristics of Three Japanese Writers

We analyzed the characteristics of the works of three famous Japanese writers, Yasushi Inoue, Yukio Mishima and Atsushi Nakajima [6].

In Japanese, several commas are used in a sentence. As our input data, we used data that examined the frequency of use of six kinds of tokens in front of the comma to determine the characteristic of the artist. The six tokens were “to”, “wa”, “de”, “toki”, “ato”, and “e”. We examined twenty-one different documents. Eight of the documents were written by Mishima, four by Inoue, and the remaining nine by Nakajima. There were six variables in the data. The classification variable was the number representing the writer. Figure 5 shows the constellation graph when no weighting factors have been assigned. Figure 6 shows the graph when the weighting factors have been included. Table 1 gives the values of the weighting factors.

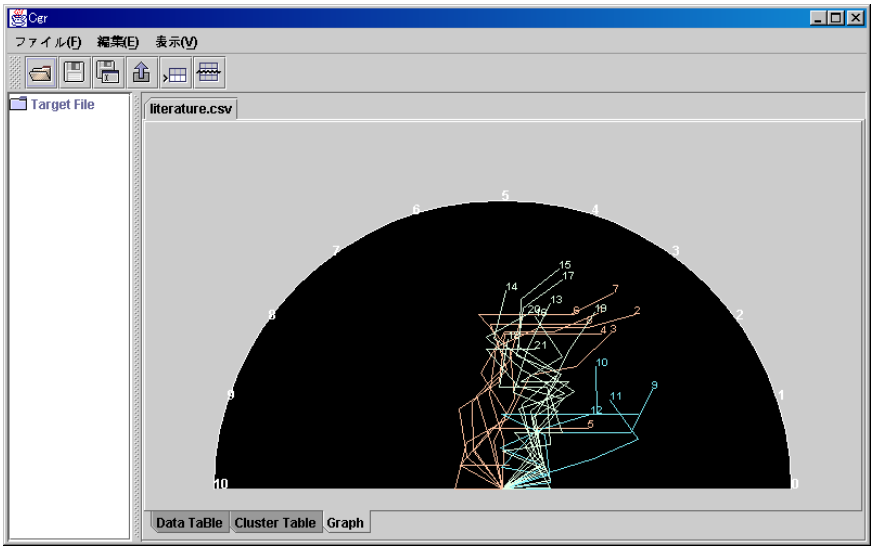

Fig. 5. The graph using the unit weighting factors.

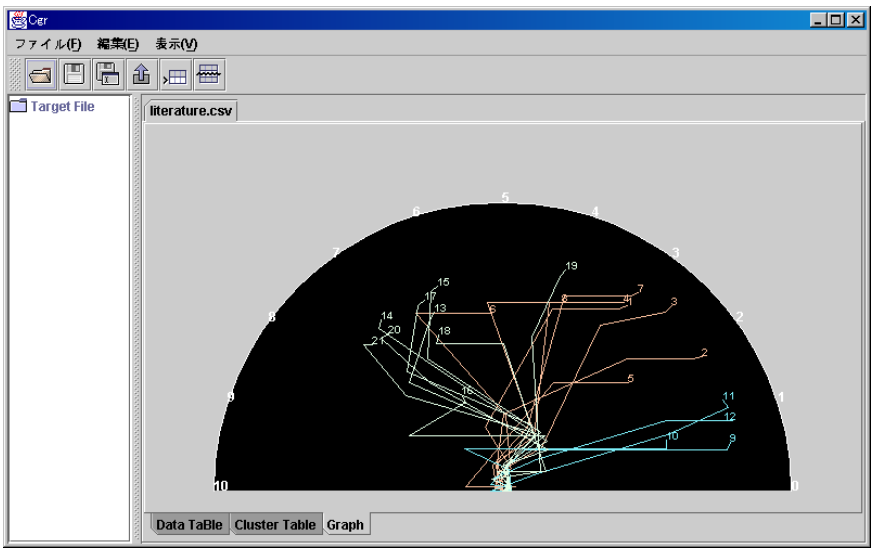

Fig. 6. The graph after incorporated the optimum weighting factors. All weighting factors are calculated so that the within-cluster variance has the minimum value and the inter-cluster variance has the maximum value.

Table 1. The weight table showing the characteristics of the writers

Writer	Weight "to"	Weight "wa"	Weight "de"	Weight "toki"	Weight "ato"	Weight "e"	The min. of J
Inoue	2.3	0.0	0.0	0.0	7.7	0.0	0.04
Mishima	0.7	4.3	1.3	0.7	3.0	0.0	0.01
Nakajima	3.7	3.7	0.0	2.3	0.3	0.3	0.13
Divides all clusters	0.3	0.7	1.7	4.7	2.3	0.3	0.85

The characteristic of the three writers can be determined from the constellation graph and the values of the weighting factors. The weight on the Table1 shows the characteristics of each writer and the value that divides all clusters on the graph. For example, if we need to distinguish only about writer "Inoue" on the graph, it is good to incorporate the weight value "to"= 2.3 and "ato"= 7.7 (the slide bar is defined from 0 to 10 on the graph). It is possible to use for the judgment of the literary works of the author not to understand, too.

3.7 An Example Analyzing Diabetes Diagnosis Data

Our next example shows the results that were obtained using diabetes checkup data of 145 people who were not overweight [7]. The data consist of five items X0 to X4:

X0: the relative weight.

X1: the blood sugar when the person became hungry.

X2: the area under the serum sugar curve when the person was given sugar when he/she became hungry; the value was recorded for three hours at 30-min intervals.

X3: the area under the serum insulin curve when after the sugar challenge recorded for three hours at 30-min intervals.

X4: the serum sugar level at equilibrium state after an intravenous injection of insulin and sugar.

The classification variable had three possible values:

1. The person is normal,
2. The person has chemical diabetes, and
3. The person has clinical diabetes.

Figure 7 shows the constellation graph incorporating no weighting factors. There is no discernable difference between the three different diagnoses. Figure 8 shows the graph when the weighting factors are included. Table 2 gives the values of the weighting factors.

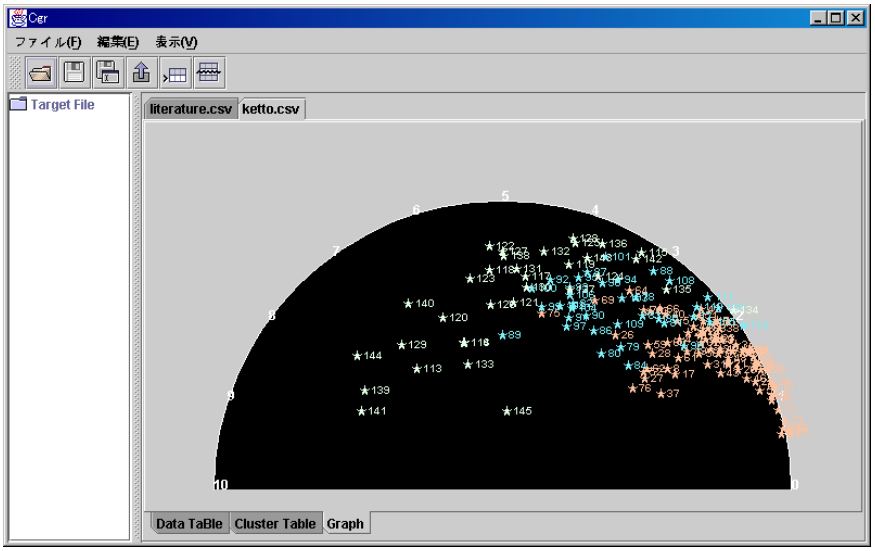

Fig. 7. The graph using the unit weighting factors.

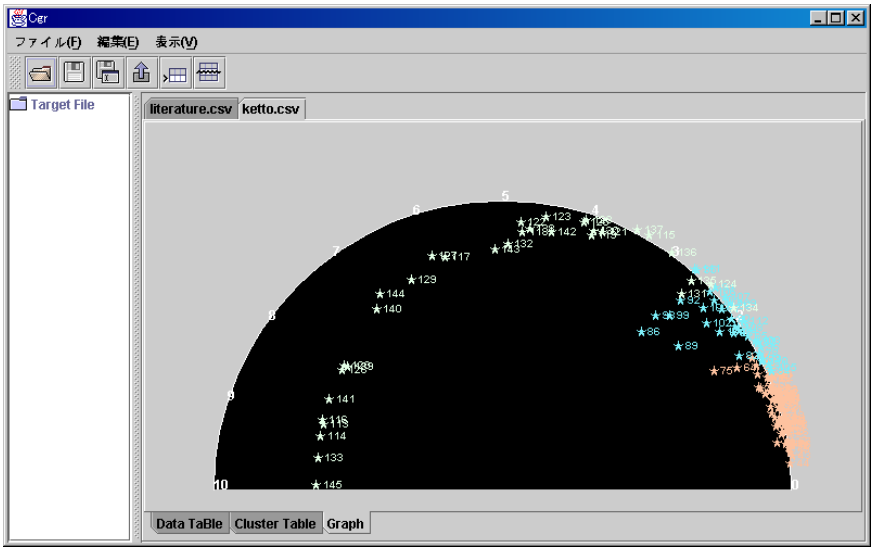

Fig. 8. The graph after incorporated the optimum weighting factors. All weighting factors are calculated so that the within-cluster variance has the minimum value and the inter-cluster variance has the maximum value.

Table 2: The weighting factors, showing the kind of the diabetes diagnosis

Daibetes	Weight X0	Weight X1	Weight X2	Weight X3	Weight X4	The min. of J
Normal	0.0	4.7	5.3	0.0	0.0	0.16
Chemical diabetes	1.0	5.7	3.3	0.0	0.0	0.23
Clinical diabetes	0.0	0.0	4.7	0.0	5.3	0.45
Divides all clusters	0.0	0.0	8.3	1.7	0.0	1.61

5 Results and Discussion

A method of constructing a constellation graph using weighting factors for the variable was presented. In addition, we have shown how an operator can select areas of the constellation graph for deletion or separate analysis at a later time. Two examples showing how our program can be used to effectively data mine a constellation graph were discussed. The most advantageous feature of our program is the provision of a means by which the operator can dynamically change the weighting factors and have these changes reflected instantly in the graph. This significantly enhances the operator's ability to cluster the data for data mining. However, visual data mining using our program is not suitable for large amounts of data due to limitations of the graphical display.

When the classification variable is categorical data, such as in our diabetes example, it is time-consuming to assign a unique value to each category when the number of categories is large or is not easily identified. Our future work will address this problem by proposing a new method for handling categorical data.

References

1. http://www.kdnuggets.com/software/visualization.html
2. Hiromi KATO: Visual Multi-Dimensional Analysis (Visualizing OLAP), IPSJ Magazine Vol.41 No.4 Apr.2000
3. Wakimoyo K., and Taguri M. : Constellation graphical method for representing multi-dimensional data. Ann. Inst. Statist.Math.30 (1997) 97-104
4. Mardia, K., and Jupp, P.: Directional Statistics John. Wiley & Sons Ltd. (1999)
5. Fisher, N.: Statistical analysis of circular data. Cambridge University Press. (1993)
6. Oyama, M., and Okada, T.: Extraction of Text Style Characteristics by Knowledge Discovery Method. K.G.Studies in Computer Science, vol.11(1996)23-36
7. Andrews, F., and Herzberg, M.: Data A Collection of Problems from Many Fields for the Student and research Worker. Springer(1985)

Appendix

Let p be the number of variables and i be the number of observations.

$$(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip}), \qquad i = 1,2,\ldots,n \tag{8}$$

Next, $f_1, f_2, \cdots, f_p$ are given as the real number functions.

$$\theta_{ij} = f_j(x_{ij}), \qquad i = 1,2,\cdots,n, \qquad j = 1,2,\cdots,p \tag{9}$$

x_{ij} is changed into an angle using condition (10).

$$0 \le f_j(x_{ij}) \le \pi \qquad f_j, \quad j = 1,2,\cdots,p \tag{10}$$

If data increases continuously, equation (11) is used.

$$f_j(x_{ij}) = \frac{x_{ij} - x_j^{(1)}}{x_j^{(2)} - x_j^{(1)}}\pi \tag{11}$$

$x_j^{(1)}, x_j^{(2)}$ are determined using (12).

$$x_j^{(1)} = \min_{1 \le i \le n} x_{ij}, \qquad x_j^{(2)} = \max_{1 \le i \le n} x_{ij}; \qquad (j = 1,2,\cdots,p) \tag{12}$$

Picture a semicircle with radius 1 and mark it with degrees.

Mark a vector $\left(\cos\theta_{ij}, \sin\theta_{ij}\right)$ corresponding to x_{ij}.

Next, multiply the vector by the weight, w_j, assigned to this vector. The vector is determined by equation (13).

$$\vec{x}_i = \sum_{j=1}^{p} \left(w_j \cos_{ij}, w_j \sin\theta_{ij}\right) \quad j = 1,2,\cdots,p \qquad (\mathbf{13})$$

Now connect these vectors.

Here, $\sum_{j=1}^{p} w_j = 1$

ξ_i: the angle of $\vec{x}_i$ and the x-axis are determined using equation (14).

$$\arg\left(\vec{x}_i\right) = \xi_i = \tan^{-1}\left(\frac{\sum_{j=1}^{p} w_j \sin\theta_{ij}}{\sum_{j=1}^{p} w_j \cos\theta_{ij}}\right) \qquad (\mathbf{14})$$

$\left|\vec{x}_i\right| = r_i$ is determined by equation (15).

$$\left|\vec{x}\right| = r_i = \sqrt{\left(\sum_{j=1}^{p} w_j \sin\theta_{ij}\right)^2 + \left(\sum_{j=1}^{p} w_j \cos\theta_{ij}\right)^2} \qquad (\mathbf{15})$$

Picture a star at the termination of $\vec{x}_i$.

By repeating the above steps for all x_{ij}, the constellation graph is constructed.