Cluster Analysis using Spherical SOM

H. Tokutaka[1], P.K. Kihato[2], K. Fujimura[2] and M. Ohkita[2]
1) SOM Japan Co-LTD, 2) Electrical and Electronic Department, Tottori University
Email: {tokutaka@somj.com, kamitazv@yahoo.co.uk, fujimura@ele.tottori-u.ac.jp,
mohkita@ele.tottori-u.ac.jp}

Keywords: Spherical SOM, Cluster Analysis, Dendrogram, and Glyph Analysis Setting

Abstract— This paper proposes a cluster analysis method of benchmark data from the Fisher's iris and the Wine recognition datasets. As a result of the numerical experiment, a clustering method using the dendrogram yielded 97 % accuracy. It is difficult to display a multi-dimensional data by the dendrogram in one dimension. The ultimate visualization is by means of 3 dimensional renditions. We conclude that the best way that a multi-dimensional dataset is visualized is by a sphere, since the phase relationship of all elements is continuous.

1 Introduction

The Self-Organizing Map (SOM) is one of the artificial neural network algorithms that were invented by Kohonen [1,2]. The SOM visualizes multi-dimensional data into two dimensions, and is widely known as a technique for data-mining, classification, and prediction in engineering, the social sciences, and so on.

Here, a new data analysis technique based on the SOM is proposed. The technique can visualize a multi-dimensional dataset on a spherical diagram and perform classification in three dimensions space.

2 Cluster analysis

Let us consider the dataset in Table 1 [2], which is a well-known example as a cluster analysis problem in a plane SOM. The relationship amongst the animal labels in Table 1 cannot be understood correctly by looking at the Table. Spherical SOM on the other hand displays the dataset such that the phase relationship can clearly be applied [3,4]. The user can express a phase relationship among the data on the spherical surface. Dendrodram is another form of data analysis. It samples the dataset and checks whether the sampled data can be grouped to similar dendrogram or not. Similar data are paired up and expressed as a new group. Further, all similar paired-up groups are then expressed as a larger group.

3 Spherical surface SOM

In a usual plane SOM, a dataset is represented on a 2-dimensional plane. In the spherical surface SOM, the nodes are arranged on a competitive layer that makes the spherical surface [4]. For example, let us take the data of Table 1. If a spherical surface SOM is used, the data can be represented on the spherical surface as shown in Fig. 1(a). In the figure, only one hemisphere of the spherical surface can be seen. The black nodes show the image of the sample and the character string on the side means the name of the sample. The other hemisphere is hidden to the viewer. Like in plane SOM, we have grey shades with dark patches indicating a wall, or break or far distance between the animals. Bright parts show a valley [5].

4 Cluster analysis using a spherical surface SOM

A dataset is presented to the competition layer of the map and then trained. A distance is computed from the sample position on the map and in this way the classification is obtained. A normal plane SOM map has edges and corners. Therefore, when the numbers of nearest-neighbor nodes on the edges and corners are compared to those at the center, we see that the former are smaller compared with the latter. In a clustering task, large errors can be generated by such distortions. Spherical SOM doesn't have such distortions, and is more suited for such cluster analysis. The details of the spherical surface SOM are explained in [4].

The following is the basic analysis procedure for the spherical SOM.

1. Dataset (e.g., Table 1) is to be presented to the learning algorithm of the spherical SOM. As a result, the spherical surface (with radius 1) map onto which a sample can be projected is initially created. When the high dimensional dataset is entered, a distortion is caused by the projection onto a lower-dimensional map. The distance among the samples in case the mapping doesn't agree with the value that was represented on the spherical surface causes distortion. Therefore, U-matrix method [1, 5] is used and a Euclidean distance among the nodes is expressed in the gray scale form. In case of mapping, the dark part means a wall and the white part means a valley as shown in Fig. 1(a).
2. Fig. 1(b) is a transformed distance mapping (Glyph) of Fig. 1(a). As a result, part of the wall and the valley are emphasized such that the three-dimensional mapping which emphasized a wall for the darkest part

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

becomes radius 1 and the brightest part 0.5. This value 0.5 can be adjusted from 0 to 1. When the brightest part of the sphere radius is equal to the darkest part (same radius 1), the structure of the map is a sphere.

3. A thicker solid line indicates the distance of the solitude between L and H in Fig. 2(a). Details of the procedure are explained in the figure caption. This procedure of computation is used even when the spherical surface is deformed. The glyph values are analyzed between 0-1 like that of procedure 2.

4. A dendrogram can be constructed when the distance among the samples is computed in procedure 3, and classification carried out. Fig. 2(b) shows the resultant dendrogram from Fig. 1(b) with the variable distances clearly expressed. Using Group-Average method non-similarity between animals can clearly be observed. The names of the animals stand uniquely in the right. For example, it can be understood that the pigeon and the hen pair up near the non- similarity value 0.35.

In order to use the spherical surface SOM, the software "blossom"[6] was utilized as a tool for the steps 1-3 due to its easy-to-use features. “blossom” can express a spherical surface in terms of a Glyph and also, in U-matrix representation of dataset in a polygonal form. The radius of the spherical surface is originally 1. When the U-matrix is expressed in a polygonal form, the default values for both biggest and smallest Euclidean distances are set as 1 and 0.5 respectively. However, using "Glyph Analysis Setting", user can change the smallest distance value freely from 0 to 1.

5 The data analysis using the database

The benchmark datasets used for the analysis are hereby described: Iris dataset of Fisher [7, 8] which is well known benchmark has a total of 150 data points arranged in the classes of 50 data points each (setosa (50), virginica (50) and vergicolor (50)). Each data point consists of the length and the width of the sepal, the length and the width of the petal, which makes it a 4-dimensional dataset. Wine data that is taken from UCI database [8]. Three kinds of wine are evaluated by the following 13 dimensions; (A) Alcohol, B) Malic acid… K) Hue, L) OD280/OD315 of diluted wines, and M) Proline). Wine 1 consists of 59, wine 2, 71, wine 3, 48 data points, thus, 178 data points in total.

6 Datasets analysis results

The effect of "Glyph Analysis Setting" was examined using the two datasets. The analysis result shown in Table-3 is the outcome of the classification obtained from ordering the final stage of the dendrogram. Tabled are the results for the grouped average, ward, flexible and centroid methods of cluster analysis. Other three namely median, nearest neighbor and furthest neighbor were given less priority due to their unstable results. The accuracy rate of group average method (GAM) seems to be highest compared to the other methods. This is mainly due to the mode of distance measurement where GAM uses the 1st power of the distance to the surface of the sphere whereas other methods use the square of the distance (virtual distance) [9].

Glyph analysis setting values can be varied from 1 to 0. When the setting is 1, then the darkest region is 1 and the brightest 0. Table 2 indicates various effects of the settings.

Table 2: Glyph setting analysis values

G. Setting	Darkest	Brightest	Sphere surface
1	1	0	Deformed-4*
0.7	1	0.3	Deformed-3
0.5	1	0.5	Deformed-2
0.2	1	0.8	Deformed-1
0	1	1	Smooth-0

* Heavily deformed

Datasets were then classified into more than three groups each having three categories of analysis depending on the setting (1, 0.5, 0). GAM in all the groupings showed the best results. It can then be concluded that GAM suits more to spherical cluster analysis than the other methods for it expresses the phase more clearly. Incidentally, in the flexible method, ($Beta$) value of the free parameter is ($-1/4 \le Beta < 0$). Here, an often-used value $Beta = -1/4$ [9] was used.

Usually, in clustering, when a dendrogram is obtained, the accuracy is judged by the fact whether the result is convenient to explain the experimental data or not. Considering the multi-dimensional dataset being projected to a single dimensional dendrogram and then being taken as conclusive is a very unstable situation that requires a buffer to ascertain the accuracy of the results. With spherical surface SOM, the result of the dendrogram and the position result of an input label on the spherical surface are compared. Referring to table 3, five mistakes can be noted in the case of G_1 of iris.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

Table 1: 16 kinds of animals and their 16 attributes [2].

		dove	hen	duck	goose	owl	hawk	eagle	fox	dog	wolf	cat	tiger	lion	horse	zebra	cow
	small	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
state	medium	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
	large	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	nocturnal	0	0	0	0	1	0	0	0.5	0	1	0.5	0.5	0	0	0	0
	2legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
possession	hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	mane	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0
	feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	stripes	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
	hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
habit	run	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
	fly	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
	swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
food life	herbivo	0.5	0.5	0.5	0.5	0	0	0	0	0	0	0	0	0	1	1	1

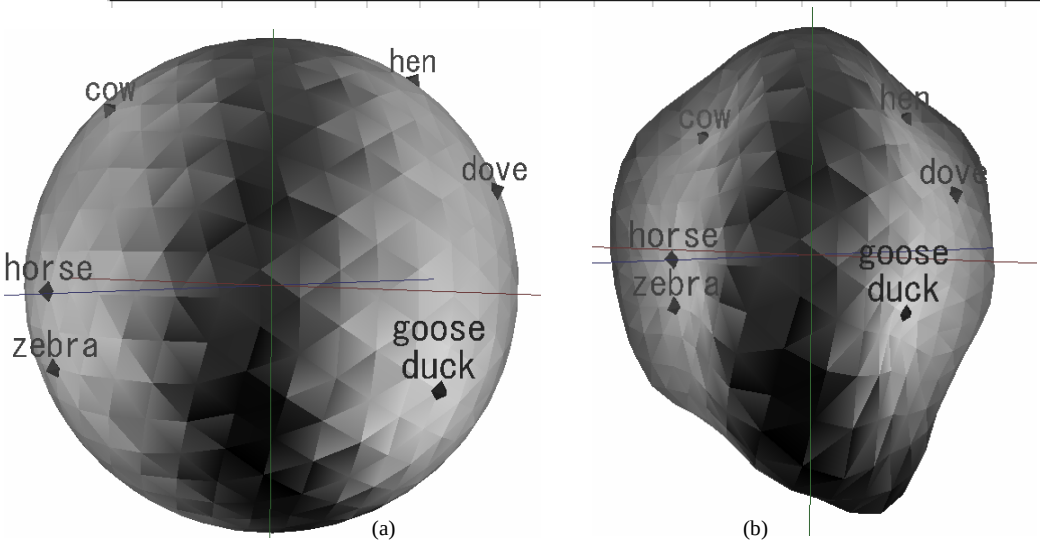

Fig. 1 (a): The result of the analysis by a spherical surface SOM for 16 kinds of animals in Table 1.
(b): The polyhedron display which follows procedure 4.

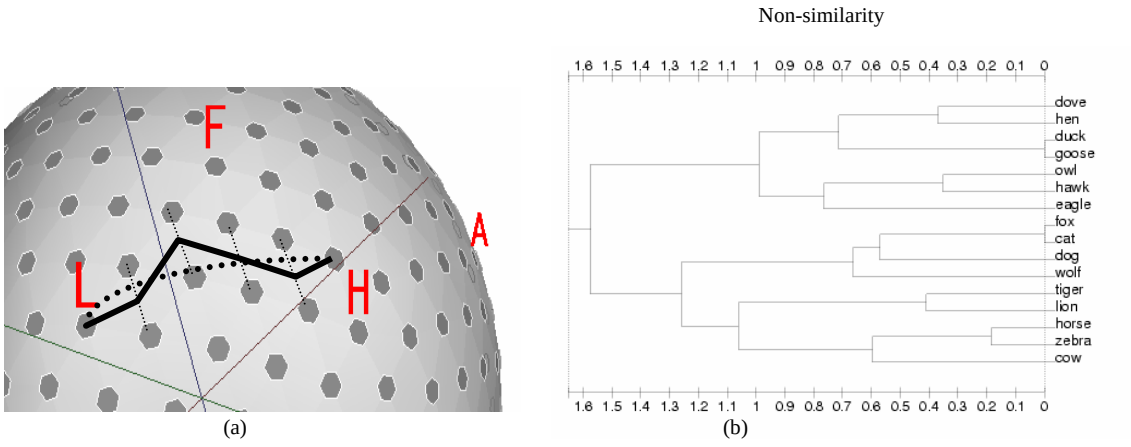

Fig. 2 (a) As for the calculation method, the dotted line with distance from the node L to H, shows the solitude. Next, the nodes that are near the solitude are chosen among the nodes that cut the solitude. The bisector of the chosen nodes is pulled and is linked by the polygonal line. The distance of this thicker solid line substitutes the distance of the solitude between L-H. (A, F are other non-related nodes). Fig. 2(b) shows the classification result obtained using information obtained from Fig. 1(b) and the distance calculation as explained in Fig. 2(a). The details analysis is as explained in procedures 1-4.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

Table 3: The results for the benchmark datasets of iris and wine recognition examined with a spherical SOM. The classification was examined from the order in the last stage of the dendrogram.

		G_1	G_0.5	G_0	W_1	W_0.5	W_0	F_1	F_0.5	F_0.5	C_1	C_0.5	C_0
iris 3 kinds	No. of Errors	5	6	17	25	23	30	5	15	14	5	X	X
4 dimensions	Accuracy (%)	97	96	89	83	85	80	97	90	91	97	X	X
wine 3 kinds	No. of Errors	6	6	11	5	13	18	5	62	62	5	5	10
13 dimensions	Accuracy (%)	97	97	94	97	93	90	97	65	65	97	97	94

G: Group average, W: Ward, F: Flexible, C: Center of gravity, X means classification is impossible.
The numbers in the 1st line show the Glyph-analysis values.
For example, G_1 means Glyph analysis value is one in Group average method.

A dendrogram and a polygon (the polyhedron, the Glyph Analysis with value 1 are shown in Fig. 3. As seen from the dendrogram, a total of five (5) misclassified items exist namely gnc_7 of virginica, ver_19, 38, 23, and 34 of vergicolor. The spherical figure in the right of Fig. 3, the polygon clutters and the position on the surface cannot be clearly seen. Therefore, Fig. 4 of the spherical surface is examined where the Glyph Analysis value is reduced to 0.

The value gnc_7, which originally belonged to the virginica group, overhangs and approaches the vergicolor group. Then, it would be incorporated into the ver_45, 41, 10 groups. If LVQ is used, the boundary should clearly be drawn, as the gnc_7 would be as the vergicolor group (refer the solid line in the figure). Similarly ver_19, 38, 23, and 34 would be grouped. When the Glyph Analysis setting is 0, a spherical surface where ver_19 is positioned on the center of the figure is shown in Fig. 5.

It clearly shows that ver_23 is very much approaching the gnc group. Also, gnc_20 is captured by ver_19 while ver_34 approaches the gnc group too. In this way, the number of the false classifications became 5. Indeed, as for the gnc_20 and ver_19, the values of the width of the sepal and the petal are almost the same and both values of the length resembles very well from the original database [7, 8]. If gnc_20 is misclassified as vergicolor, ver_19, 38, 23, and 34, can be understood to belong to ver (vergicolor) from Fig. 5. Thus, the boundary of ver (vergicolor) and gnc (verginica) can be distinguished from the learning result drawn onto the spherical surface. After all, if only gnc_20 is considered captured by vergicolor, it can be understood that gnc_20 was misclassified. Then, it is possible to say that the correct percentage becomes 99 %, thus 149/150.

In the example of iris, a Group-average method and Glyph Analysis setting of 1.0 were used for the analysis. As shown in Table 3 and Fig. 3, there were no-misclassifications in setosa; virginica had one while virgicolor had four making a total of five over the 150 iris stocks. Therefore, an accuracy of 97 % in the data analysis was realized.

Next, the unknown data X_7 with a group location data 5.1, 2.8, 4.8, and 1.76 of the sepal length, sepal width, petal length, and the petal width with each other was prepared. After learning with Fig. 3, the X_7 label was pasted as shown in Fig. 4. It is noted that X_7 is near gnc_7 and from the boundary of the solid line, it can be understood that X_7 would belong to gnc (virginica) group.

Further the number of iterations was increased to 500 for the training of spherical SOM. Data value gnc_20 and ver_19 can vividly be seen to be separate as shown in Fig. 6. The solid line is the boundary between the gnc group and the ver group. The ver_23 and ver_34 can be seen to overhang to the gnc group. If the gnc_20 can be incorporated into the ver group as misclassification, it is possible to draw the boundary between the gnc group and the ver group smoothly as shown by the dotted line.

In the example of wine recognition dataset, when the Glyph Analysis Setting was set as 1.0 using the Group-average method, Table 3, indicated a misclassified number of 6. Here, two of wine_2 were misclassified as wine_1 and four as wine_3. When a boundary was drawn like in the previous example, these misclassified items were possible to be admitted to the group of wine_2. Incidentally, there were no misclassifications in the groups of wine_1 and wine_3. This time, the optimization of Glyph Analysis Setting was only examined with increased learning time (default learning time of label number×50). Besides the optimization of the learning rate factor, the neighborhood function can also be optimized to enhance the classifications.

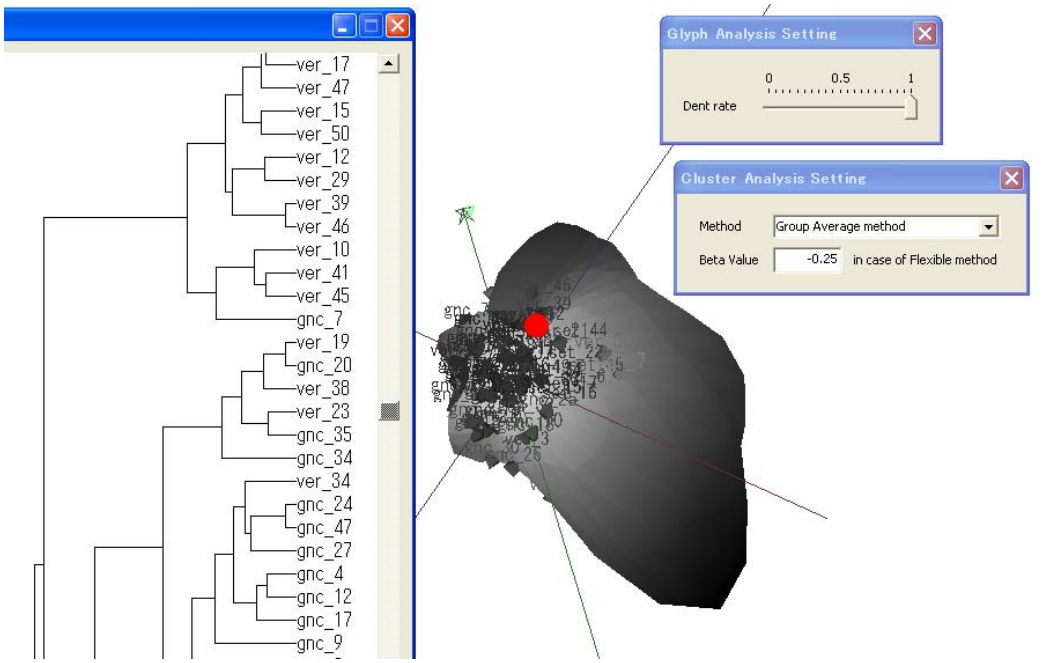

Fig. 3 Spherical surface which was transformed at Glyph Analysis value 1 and the part of the dendrogram (Group-average method) determined from the distance calculation on the transformed surface. (The misclassified part: the gnc_7 of virginica as ver(vergicolor) and also, ver_19, 38, 23, and 34 (vergicolor), as gnc(virginica) group.

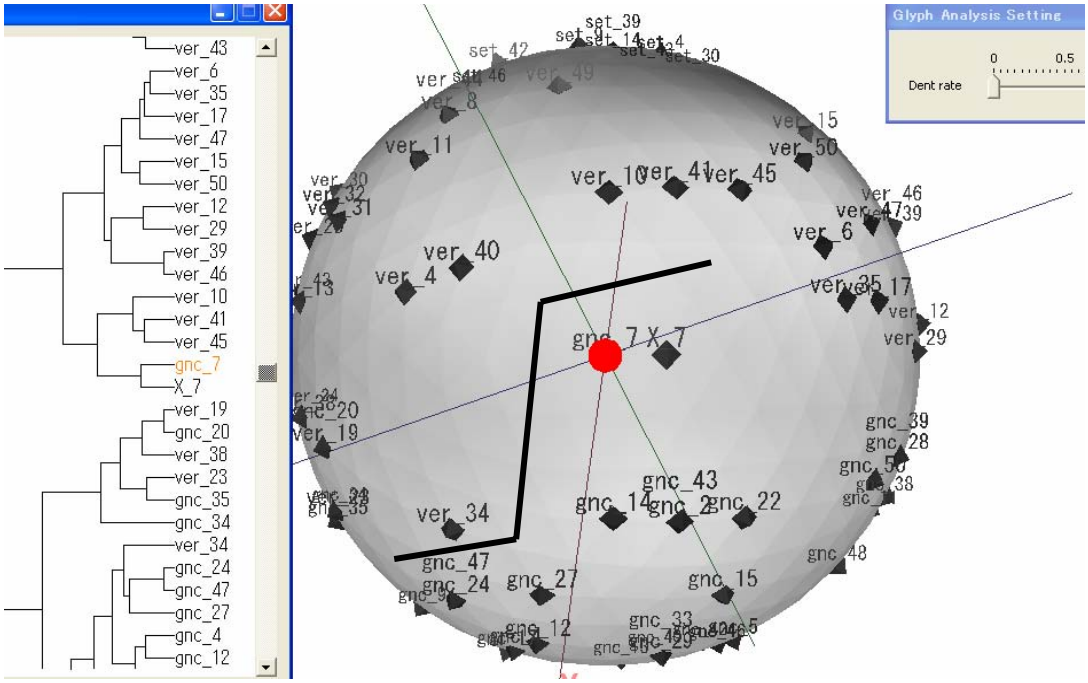

Fig.4 In order to see the position of the misclassified gnc_7 (the center of the surface), the Glyph Analysis value was reduced to 0. The dendrogram in the left side is the same as the one of Fig. 3 (Glyph Analysis value 1). The boundary of the solid line was artificially drawn. X_7 is an unconfirmed data point. From the figure, it can be found as belonging to the gnc group.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

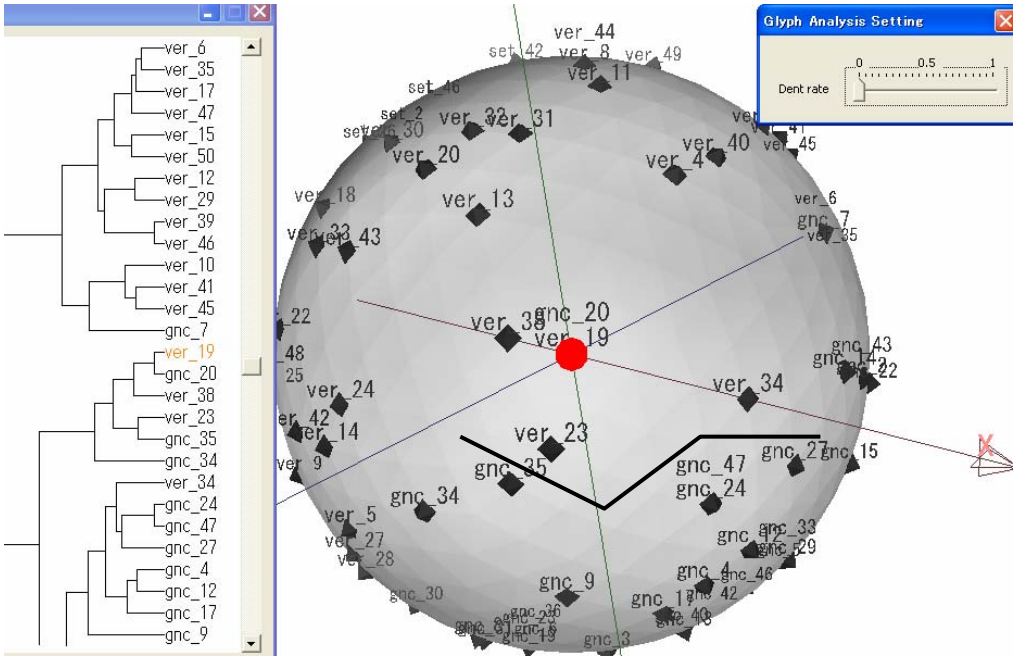

Fig. 5 In order to see the misclassified positions of ver_19, 38, 23, and 34, the Glyph Analysis value was reduced to 0 and there, the ver_19 data point is in the center. If the gnc_20 misclassified the ver group, the boundary would move between ver_23 and gnc_35 and also between ver_34 and gnc_27. The solid boundary line was artificially drawn.

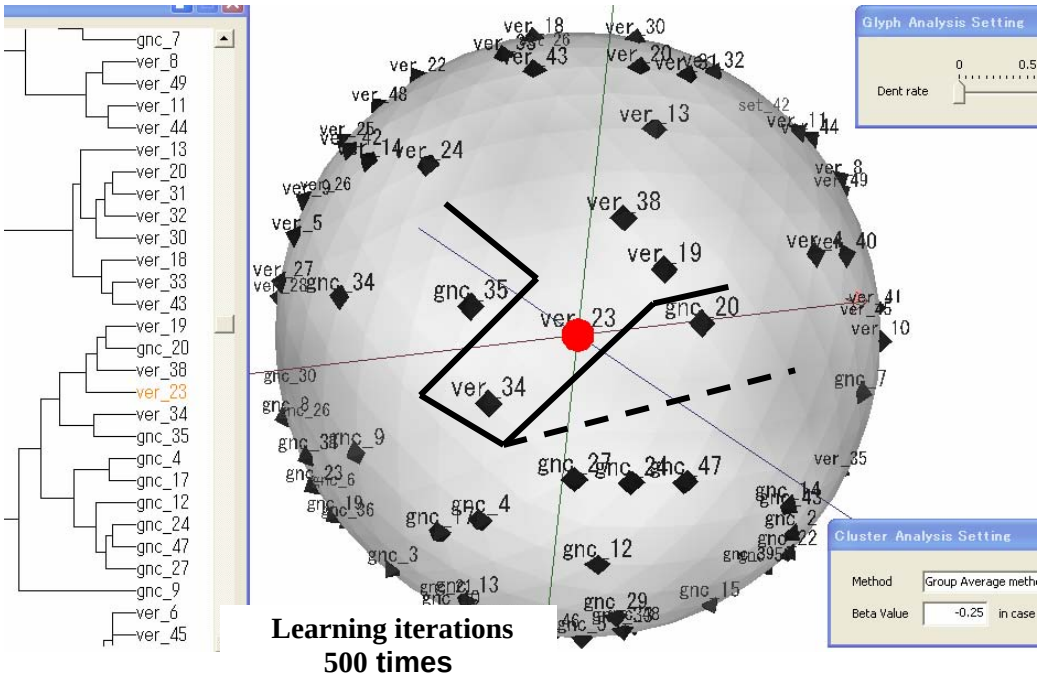

Fig. 6 Item ver_19 and gnc_20 in Fig. 5 were separated after the learning time was increased to 500 epochs. The solid line is the boundary between ver and gnc. The dotted line is the boundary when the gnc_20 was permitted to belong to the ver group.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

7 Conclusion

Here, the classification method using the spherical surface SOM was proposed. With this technique, the sample set can be visualized as a three-dimensional figure. Also, the sample set can be expressed as the dendrogram for classification purposes. Using the iris dataset of Fisher [7, 8] and the wine recognition dataset of UCI [8] as our benchmarks, the spherical surface SOM approach was evaluated. As a result, a maximum of 97 % correct classifications was obtained in the iris dataset as well as for the wine recognition dataset. Also, those misclassifications that are contained in the dendrogram were examined minutely on the three-dimensional spherical surface diagram where they are projected. From the analysis, if gnc_20 is permitted to belong to ver group, then from Fig. 5, the correct percentage becomes actually 99 %. Moreover, ver_19 and gnc_20 are distinctively separate when the learning iterations are increased to 500 times. The correct percentage of the classification could reach 100 % from the new solid boundary of Fig. 6. Also, in the example of wine, the misclassified 6 labels were corrected from the positions on the spherical surface. In this case, the correct percentage of classification becomes 100 %.

Therefore, in conjunction with the visualization of a spherical surface, a dendrogram can be used to give more details as illustrated by the examples given. The proposed approach of more effectively using the dendrogram in conjunction with other cluster analysis methods increases the reliability of the analysis. This differs fundamentally from the conventional way of displaying results on the dendrogram conclusively.

In other words, when unclassified data of iris and wine recognition sets are examined, it is possible to judge easily to which group they belong with the positions on the spherical surface of the unclassified data because the boundaries are already drawn in Figs. 4-6.

It is worth noting, that it is difficult to organize and display multidimensional data by the dendrogram in a single dimension. Hence, it is better to display the data in two or even three dimensions. Ultimate visualization would be realized by using a three-dimensional display and a sphere is the best choice since it can express a smooth phase relationship. In the example of the iris, ver_19 and gnc_20 can be separated by increasing the learning times. New boundaries were drawn on the spherical surface and the correct percentage of the classification reached 100 %. By combining the one dimensional dendrogram and the three dimensional spherical surface SOM analysis, the correct percentage of the classification could reach 100 % for the 2 benchmark problems iris [7,8], and wine recognition [8].

References

[1] T. Kohonen, Self-Organizing Maps, Springer Series in Information Sciences, Volume 30, 2001.

[2] H. Tokutaka, S. Kishida, and K. Fujimura, Self-Organizing Maps and Applications -2dimensional visualization of multi-dimensional data in Japanese, Kaibundo Publishing Co. Ltd., 1999

[3] H. Ritter, Self-Organizing Maps on non-euclidean Spaces, Kohonen Maps, Editors, E. Oja, and S. Kaski, Elsevier, pp.95-110, 1999.

[4] D. Nakatsuka and M. Oyabu, "Application of Spherical SOM in Clustering", Proceedings of Workshop on Self-Organizing Maps (WSOM'03), pp. 203-207, 2003.

[5] A. Ultsch, G. Guimaraes, D. Korus and H. Li, Knowledge Extraction from Artificial Neural Networks and Applications, Proc. TAT/WTC93, Springer, pp.194-203, 1993.

[6] http://www.somj.com/.

[7] R. A. Fisher, The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics, Vol.7, pp.179-188, 1936.

[8] http://www.ics.uci.edu/~mlearn/databases/

[9] Y.Tanaka and T.Tarui, The Stastistical Analysis Handbook - the multi-variate analysis (in Japanese), Kyoritsu Publishing Co-Ltd, p.139, 2003.