CHAPTER 2
CLUSTERING ANALYSIS WITH UNBIASED METHODS TO DIFFERENTIATE CERVIX CANCER PATIENTS AND OVARIAN CANCER PATIENTS FROM NORMAL PERSONS USING PROTEIN CONTENT OF SPECIFIC PROTEOGLYCAN OBTAINED FROM FLOW INJECTION SYSTEM

2.1 Introduction

Unsupervised pattern recognition is widely used to find hidden interrelationships in data. Groups of samples can be expressed without a prior knowledge [1, 2]. Conception of unsupervised pattern recognition can be applied in 3 algorithms: cluster analysis, eigenvector methods and neural networks [3]. *K*-means clustering and hierarchical clustering analysis (HCA) is mostly used to discriminate signals of samples in many fields e.g. food industry [4-8], pharmacy[1, 9-11] and health science and application [12-14].

Flow injection system with mini-immunoaffinity chromatographic column for chondroitin sulfateproteoglycans assay was an effective method to screen various types of cancers [15]. Value of "relative amount of protein content in specific proteoglycan per 100 mg total protein" (referring as "WF6".) was used to screen 5 types of cancers. The details of experimental results were reported in [15] and [16]. Although the successful screening can be for ovarian cancer and cervix cancer screening, some WF6 values are unclear for the diagnostic purposes.

In this work, we demonstrate the application of chemometrics to the data previously studied to confirm the screening using WF6 for ovarian and cervix cancer samples using FIA procedure [1], by applying unsupervised pattern recognition,

namely, *k*-means clustering and HCA. This involved 2 separate datasets. (i.e. the sets for ovarian and cervix cancer samples.)

2.2 Chemometrics methodology

Chemometrics methods can be used as investigation tool of data. Clustering of univariate data is done in condition of unbiased data. Box plot is one of outlier detection to get rid the extreme and outlier value in data. In this study unsupervised pattern recognition methods obtained HCA and *k*-means clustering are applied to express distribution and inner correlation of data.

2.2.1 Box plot [17-19]

Box plot is graphical box of overall information in data in form of whiskers rectangle. Whisker included median and interquartile range. From the whisker, median of data is shown as line in the whisker and the line above and below from median as1.5 times the interquartile range. Outlier values are the objects which plotted outside the whisker. The outliers are left out to get rid bias value(s) that affect(s) descriptive statistics of data.

2.2.2 Clustering methods

Clustering methods can be performed using 2 algorithms; hierarchical and partition clustering [20]. In this study, HCA and *k*-means clustering were employed.

2.2.2.1 Hierarchical Clustering Analysis [8, 21, 22]

HCA involves agglomerative process to classify WF6 values of samples in the data and shown as dendrogram. Square Euclidean distance of all pairs of WF6 values are calculated to measure similarity of all WF6 values in the data. The closest distance WF6 pair is joined first to combine to 1 cluster. The distances of the

cluster and the rest of data are recalculated. The closest WF6 values and the cluster joining together in the first time are combined by using between group linkage to make a bigger cluster. Agglomerative hierarchical process is performed until all of WF6 values are grouped into a single group and expressed by dendrogram. Characteristics of groups in the data can be noticed from rescaled distance clusters combined in dendrogram.

2.2.2.2 *K*-means clustering [23]

K-means clustering is partition method assigning WF6 values into non-overlapping groups. Number of groups will be determined by user. Euclidian distance of each WF6 value of sample to the center of cluster was iteratively calculated. The maximum of iterations is fixed as 100 with convergence criterion is 0. The centers are selected when the least of sum of square of Euclidian distance from all members of the cluster to cluster center is found.

2.3 Experimental

Chemometrics processes performed in this work are illustrated in Figure 1. It involves two-steps processes for clustering of WF6 values employing 2 methods; HCA and *k*-means clustering. The clustering results are compared with the hospital records.

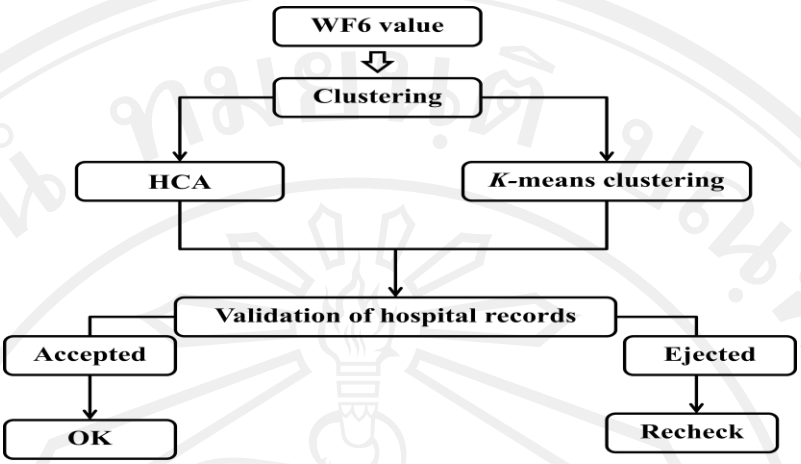

Figure 2.1 Chemometrics process of clustering of WF6

2.3.1 The data [15]

Amount of chondroitin sulfate proteoglycans can be estimated by using "relative amount of protein content in specific proteoglycan per 100 mg total protein" which was calculated from signal obtained from flow injection system with mini-immunoaffinity chromatographic column. Detail of the studies was published in reference [15]. Specific proteoglycans can be done by binding of WF6 (which served as antibody).In this work, "relative amount of protein content in specific proteoglycan per 100 mg total protein", referring as WF6, was used to screen cancer patients. From the reference [15], it was found that cervix cancer and ovarian cancer samples indicate trends to separate the patients from healthy people, although some cases were unclear. Hospital records of each group refer to 25, 12, and 14 for healthy people, cervix cancer patients, and ovarian cancer patients respectively ("CC", "OC" and "N" referring to the cervix cancer, ovarian cancer and healthy cases, respectively.).

In this work, chemometrics methods were used to confirm the screening of cervix and ovarian cancer from healthy people .According to the previous report [1],

the results involved 2 databases; set I concerning cervix cancer cases while set II concerning ovarian cancer cases. Treating to be normal or case sensitive was the according the hospital records.

Clustering methods, with unsupervised pattern recognition, *k*-means clustering and HCA were employed to divide the signals into 2 groups with expectation to be cancer and healthy cases; N (negative) and CC (cervix cancer) and OC (ovarian cancer) for datasets I and II, respectively. The results of clustering were used to reflect distribution of WF6 inside the datasets. The misclassified values from the clustering were found out and needed more study to confirm grouping of samples.

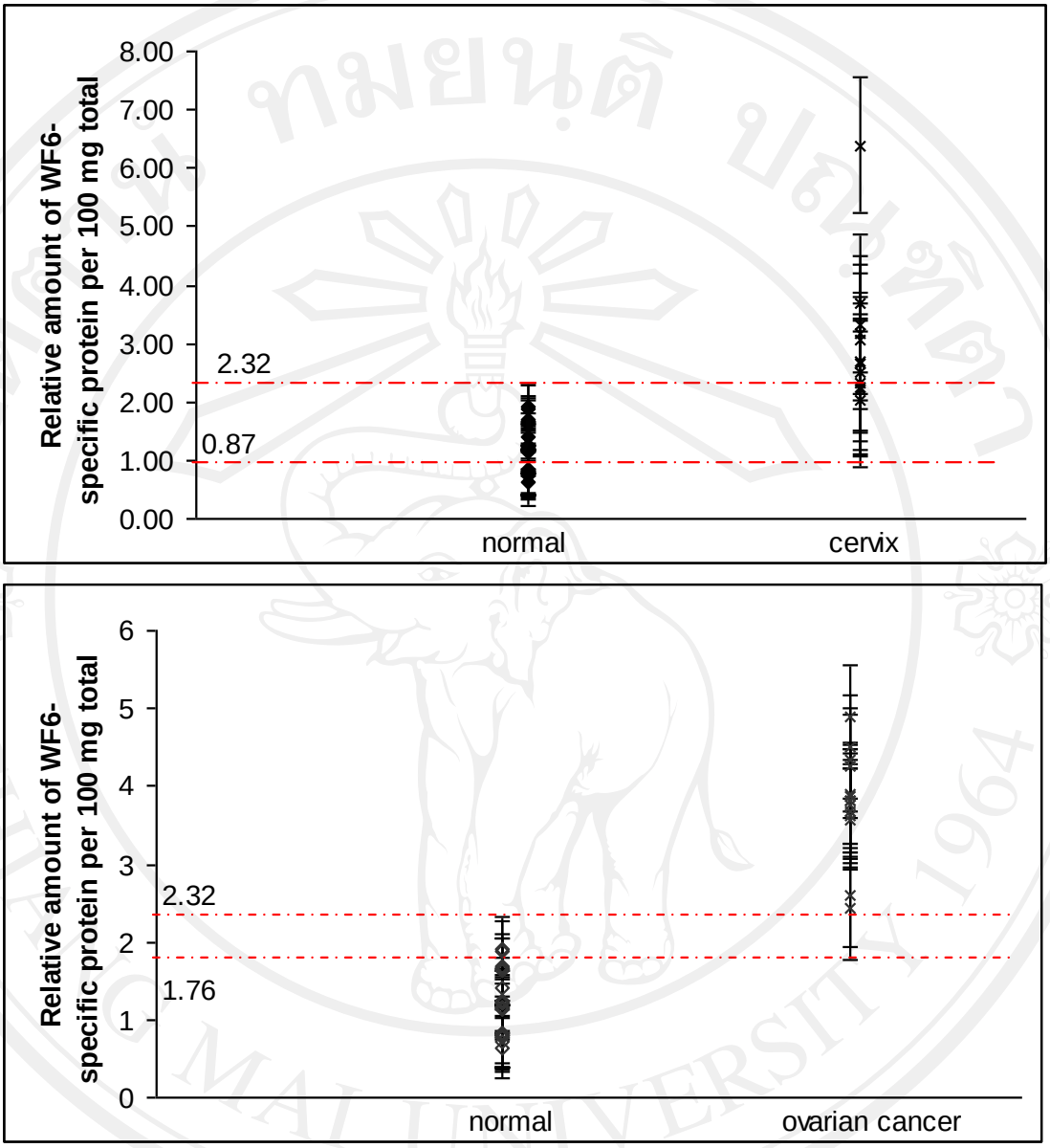

Figure 2.2 Relative amount of WF6-specific protein per 100mg total protein ±SD of normal-cervix cancer cases (top) (database I) and normal-ovarian cancer cases (database II) (bottom) [1]

2.3.2 Apparatus

Chemometrics methods used in this work were performed by all of softwares running via an AMD Athlon™ 64X2 Dual Core Processor 3800+ 2.20 GHz., 2.00 GB of RAM Physical Address Extension.

2.3.3 Chemometrics processing

2.3.3.1 Data observation

Box-plot test was used to survey outlier samples in the database. The outlier cases were removed from the database before performing clustering process to get rid bias of outlier.

2.3.3.2 Samples groups identification by *k*-means and HCA clustering

Groups of samples were identified by 2 clustering methods; *k*-means clustering and HCA and were called G.k-means andG.HCA, respectively. The validation of clustering results was done by comparing with hospital record. Validation terms used to explain clustering results were sensitivity, specificity and probability positive. Definitions of those 3 terms [15] are: SV or sensitivity is number of ratio of diseased patients with positive test per number of diseased patients; SP or specificity is number of ratio of nondiseased patients with negative test per number of nondiseased patients; Prob. positive is probability of diseased patient if the test is positive.

2.4 Results and discussion

2.4.1 Dataset I: the cervix cancer-normal cases

2.4.1.1 Data observation

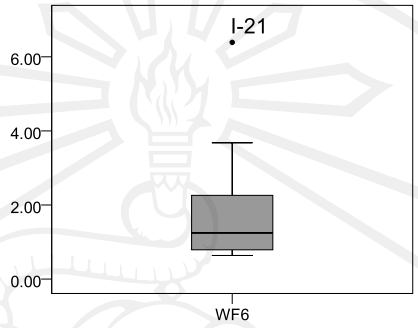

Figure 2.3 Box test of database I. normal-cervix cancer

WF6 values of dataset I were checked by box plot as shown in Figure 3. It was found that distribution of data was weighted to the low WF6 values. From the box plot, the sample number I-21 was defined as outlier due to too high WF6 value (6.39). The WF6 value of sample number I-21 was pulled out from the dataset before performing clustering process.WF6 value of sample number I-21 was very clear to identify as “CC”.

2.4.1.2 Samples groups identification by *k*-means clustering and HCA clustering

Hidden groups of WF6 values within dataset I were expressed by dendrogram in Figure 4. The samples were separated to 2 groups. Distribution of WF6 values can be also noticed from the dendrogram that inside N and CC group, there were1 and 2 groups, with population of 20 and 16, respectively. The distribution in CC group indicated that there were 2 levels of illness by this cluster members. As there were 2 groups, so the number 2 was assigned as number of group in *k*-means clustering.

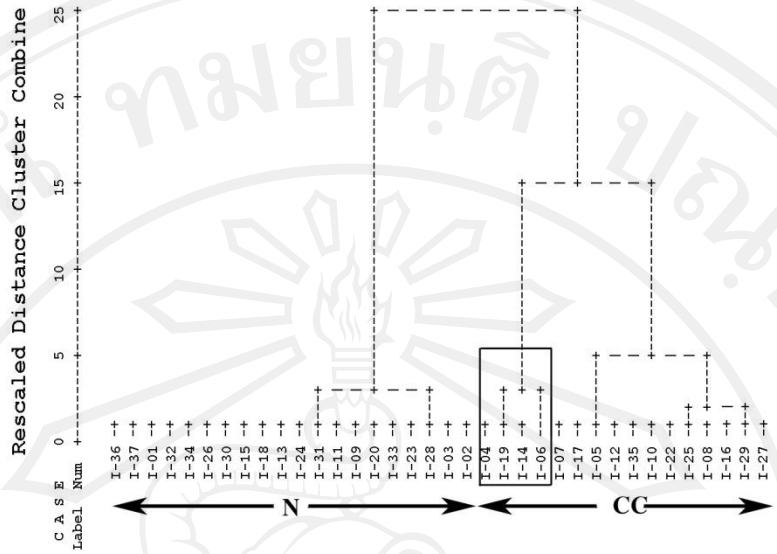

Figure 2.4 Dendrogram of database I (N was normal and CC was cervix cancer)

For HCA, the results were compared to that of the hospital records, it was found that sensitivity, specificity and probability positive were 1.00, 0.92 and 0.85, respectively. The misclassified WF6 values refer to the samples: I-5, I-7, I-12, I-17 and I-35. From Figure 4, CC group exhibits 3 mini groups. The misclassified WF6 values were clustered in the same group (see the blog in Figure 4). WF6 values of I-12 and I-35 were taken into CC group as I-12 and I-35 being closer to the most WF6 values of CC group. WF6 values of sample numbers I-5, I-7 and I-17 were closer to I-12 and I-35 more than WF6 values in the N group so those 5 WF6 values were grouped into CC.

Table 2.1 Clustering results of database I from *k*-means clustering and HCA and confirmation of samples by comparison of both methods

#	WF6	HCA	*k*-means clustering	Record	Reference (ROC)
I-01	0.79	N	N	N	N
I-02	1.41	N	N	N	N
I-03	1.25	N	N	N	N
I-04	3.18	CC	CC	CC	CC
I-05*	1.69	CC	N	N	N
I-06	3.68	CC	CC	CC	CC
I-07*	1.65	CC	N	N	N
I-08	2.34	CC	CC	CC	CC
I-09	0.64	N	N	N	N
I-10	2.04	CC	CC	CC	CC
I-11**	0.73	N	N	N	N
I-12**	1.92	CC	CC	N	N
I-13	0.79	N	N	N	N
I-14	3.33	CC	CC	CC	CC
I-15	0.79	N	N	N	N
I-16	2.64	CC	CC	CC	CC
I-17*	1.64	CC	N	N	N
I-18	0.79	N	N	N	N
I-19	3.05	CC	CC	CC	CC
I-20	1.14	N	N	N	N
I-21***	6.39	outlier	outlier	CC	CC
I-22	2.26	CC	CC	CC	CC
I-23	1.22	N	N	N	N
I-24	0.77	N	N	N	N
I-25	2.23	CC	CC	CC	CC
I-26	0.79	N	N	N	N
I-27	2.51	CC	CC	CC	CC
I-28	1.22	N	N	N	N
I-29	2.69	CC	CC	CC	CC
I-30	0.79	N	N	N	N
I-31	0.84	N	N	N	N
I-32	0.79	N	N	N	N
I-33	1.17	N	N	N	N
I-34	0.79	N	N	N	N
I-35**	1.87	CC	CC	N	N
I-36	0.79	N	N	N	N
I-37	0.79	N	N	N	N

Table 2.2 Summary of clustering of WF6 values in dataset I by *k*-means clustering

Hospital records		*k*-means clustering	
N	25	N	23
		CC	2
CC	11	N	0
		CC	11
SV		1.00	
SP		0.92	
prob.		0.85	

Table 2.3 Summary of clustering of WF6 values in dataset I by HCA

Hospital records		HCA	
N	25	N	20
		CC	5
CC	11	N	0
		CC	11
SV		1.00	
SP		0.80	
prob.		0.69	

K-means clustering grouped WF6 values in the dataset I into N and CC of 23 and 13 members, with cluster centers of N and CC of dataset I being 1.01and 2.60 respectively. From Table 2, the results obtained by *k*-means clustering were compared to the hospital records, it was indicated that, for CC, they agreed with that of the hospital records. So were that obtained by HCA. For the 25 N -members of the hospital records, the results by *k*-means clustering were 23members in the N group and 2members in the CC group, while the results obtained by HCA being N and CC for 20and 5, respectively. The distribution of WF6 values in dataset I could be explained by SV, SP and probability positive. From the validation dataset I; SV, SP and probability positive were found to be 1, 0.92 and 0.85, respectively.

The clustering results by HCA and *k*-means clustering showed that in the dataset I, the data of I-5, I-7, I-12, I-17 and I-35 needed further investigated.

2.4.2 Dataset II: the ovarian cancer-normal cases

2.4.2.1 Data observation

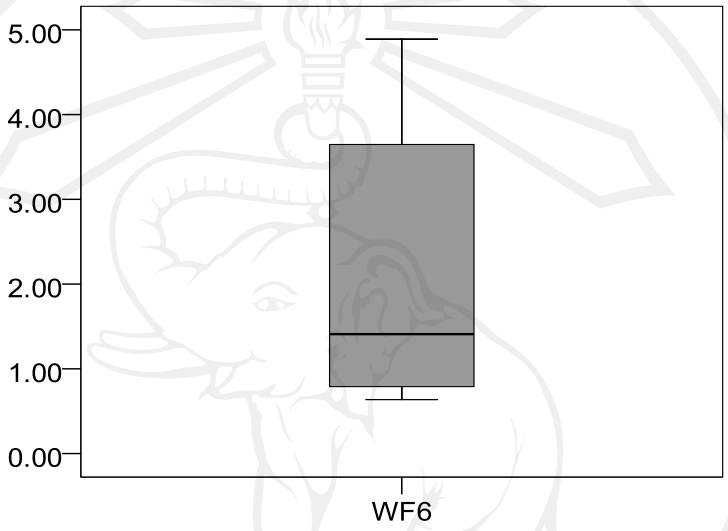

Figure 2.5 Box test of database II: normal-ovarian cancer

In figure 2.5, distribution of samples in dataset II was little weighted into low WF6 but all data still within the red box and did not show outlier case. It was shown that all samples in dataset II can be used to study.

2.4.2.2 Samples groups identification by *k*-means clustering and HCA clustering

When HCA was used to cluster data, the dendrogram shows 2 clearly groups of N and OC (Figure2. 6).

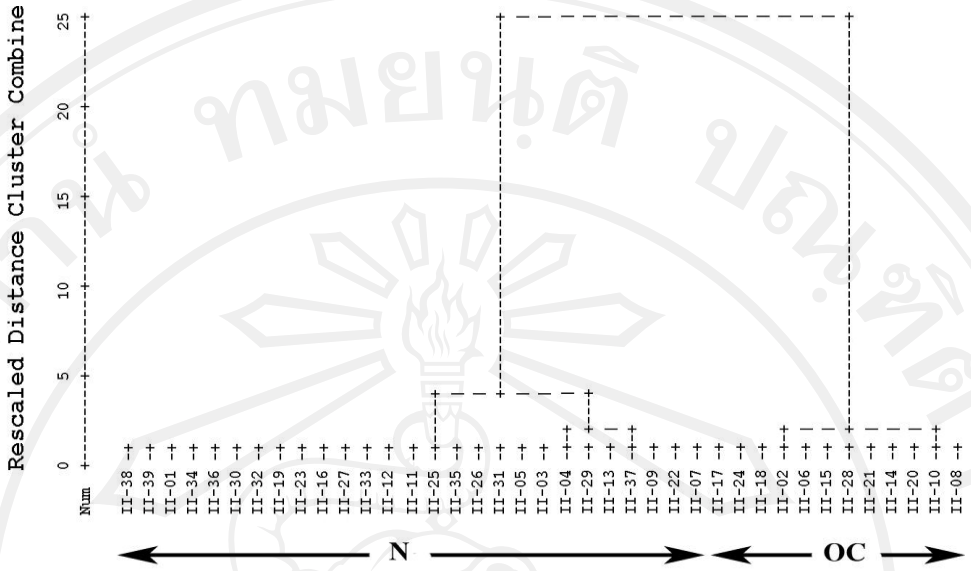

Figure 2.6 Dendrogram of database II (N was normal and OC was ovarian cancer)

From HCA, the dataset was classified into 2 groups, so 2 was number which was assigned to be K number in *k*-means clustering. Centers of the clusters of N and OC were 1.19 and 4.00, respectively. From Table 6, all WF6 values of the samples in the dataset II can be clustered by *k*-means clustering and HCA with no different results. Groups of N and OC were clearly discriminated so the results of clustering were not affected from method of clustering. Distribution of WF6 in N group was clearly discriminated so specificity and probability positive of dataset I were 1.00 and 1.00, respectively. WF6 values of OC were clustered into 2 sets, 12 WF6 values were correctly classified while 2 WF6 values of OC were misclassified. WF6 values of sample numbers II-4andII-29 were misclassified and led to sensitivity 0.86because their WF6 values were closer to center of OC group more than the center of N group .

From clustering results, sample numbers II-4 and II-29 needed other methods to confirm the results.

Table 2.4 Clustering result of database II from *k*-means clustering and HCA and confirmation of samples by comparison of both methods

#	WF6	HCA	*k*-means clustering	Record	Reference (ROC)
II-01	0.79	N	N	N	N
II-02	3.82	OC	OC	OC	OC
II-03	1.41	N	N	N	N
II-04*	2.59	N	N	OC	OC
II-05	1.25	N	N	N	N
II-06	3.81	OC	OC	OC	OC
II-07	1.69	N	N	N	N
II-08	4.89	OC	OC	OC	OC
II-09	1.65	N	N	N	N
II-10	4.51	OC	OC	OC	OC
II-11	0.64	N	N	N	N
II-12	0.73	N	N	N	N
II-13	1.92	N	N	N	N
II-14	4.25	OC	OC	OC	OC
II-15	3.86	OC	OC	OC	OC
II-16	0.79	N	N	N	N
II-17	3.58	OC	OC	OC	OC
II-18	3.68	OC	OC	OC	OC
II-19	0.79	N	N	N	N
II-20	4.35	OC	OC	OC	OC
II-21	3.91	OC	OC	OC	OC
II-22	1.64	N	N	N	N
II-23	0.79	N	N	N	N
II-24	3.62	OC	OC	OC	OC
II-25	1.14	N	N	N	N
II-26	1.22	N	N	N	N
II-27	0.77	N	N	N	N
II-28	3.77	OC	OC	OC	OC
II-29	2.42	N	N	OC	OC
II-30	0.79	N	N	N	N
II-31	1.22	N	N	N	N
II-32	0.79	N	N	N	N
II-33	0.84	N	N	N	N
II-34	0.79	N	N	N	N
II-35	1.17	N	N	N	N
II-36	0.79	N	N	N	N
II-37	1.87	N	N	N	N
II-38	0.79	N	N	N	N
II-39	0.79	N	N	N	N

Table 2.5 Summary of clustering of WF6 values in dataset II by HCA and *k*-means clustering

Hospital records		*k*-means clustering and HCA	
N	25	N	25
		OC	0
OC	14	N	2
		OC	12
SV		0.86	
SP		1.00	
prob.		1.00	

2.5 Conclusion

Unsupervised pattern recognition, *k*-means clustering and HCA can be used as tools to investigate the distribution of WF6 values in the previous studied results: datasets I and II for ovarian and cervix cancer cases. The differentiation confirms the previous report [1], although there are a few cases that need further investigation in detail, including to trace the validation of the patients' records.

2.6 References

1 K. C. Weber, A. B. F. D. Silva, *European Journal of Medicinal Chemistry* **2008**, *43*, 364-372.

2 A. P. Fernandes, M. C. Santos, S. G. Lemos, M. M. C. Ferreira, A. R. A. Nogueira, J. A. Nóbrega, *Spectrochimica Acta Part B: Atomic Spectroscopy* **2005**, *60*, 717-724.

3 H. Philip K, *Analytica Chimica Acta* **2003**, *500*, 365-377.

4 E. Marengo, M. Aceto, V. Maurino, *Journal of Chromatography A* **2002**, *943*, 123-137.

5 C. Sola-Larrañaga, I. Navarro-Blasco, *Analytica Chimica Acta* **2006**, *555*, 354-363.

6 N. Niemenak, C. Rohsius, S. Elwers, D. O. Ndoumou, R. Lieberei, *J. Food Compos. Anal.* **2006**, *19*, 612-619.

7 W.-J. Wang, Y.-X. Tan, J.-H. Jiang, J.-Z. Lu, G.-L. Shen, R.-Q. Yu, *Chemometrics and Intelligent Laboratory Systems* **2004**, *72*, 1-8.

8 A. Akbay, A. Elhan, C. Özcan, S. Demirtas, *Medical Hypotheses* **2000**, *55*, 147-154.

9 F. A. Molfetta, A. T. Bruni, K. M. Honório, A. B. F. da Silva, *European Journal of Medicinal Chemistry* **2005**, *40*, 329-338.

10 J. Ruiz-Jiménez, F. Priego-Capote, J. García-Olmo, M. D. L. d. Castro, *Analytica Chimica Acta* **2004**, *525*, 159-169.

11 J. Souza Jr, R. H. de Almeida Santos, M. M. C. Ferreira, F. A. Molfetta, A. J. Camargo, K. Maria Honório, A. B. F. da Silva, *European Journal of Medicinal Chemistry* **2003**, *38*, 929-938.

12 T. Yukio, *Chemometrics and Intelligent Laboratory Systems* **1999**, *49*, 105-115.

13 K. Deb, A. Raji Reddy, *Biosystems* **2003**, *72*, 111-129.

14 D. Milde, J. Macháček, V. Stužka, *Chemical Papers* **2007**, *61*, 348-352.

15 S. K. Hartwell, K. Pathanon, D. Fongmoon, P. Kongtawelert, K. Grudpan, *Anal Bioanal Chem* **2007**, *388*, 1839-1846.

16 K. Pathanon, PhD. chiang-mai university, Thailand, **2003**.

17 K. Melody Y, *Decision Support Systems* **2003**, *35*, 441-454.

18 E. W. Steyerberg, F. E. Harrell Jr, G. J. J. M. Borsboom, M. J. C. Eijkemans, Y. Vergouwe, J. D. F. Habbema, *Journal of Clinical Epidemiology* **2001**, *54*, 774-781.

19 M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, B. Walczak, *Chemometrics and Intelligent Laboratory Systems* **2007**, *85*, 203-219.

20 A. Ahmad, L. Dey, *Data & Knowledge Engineering* **2007**, *63*, 503-527.

21 F. Gong, B.-T. Wang, Y.-S. Fung, F.-T. Chau, *Atmospheric Environment* **2005**, *39*, 6388-6397.

22 R. E. Abdel-Halim, R. E. Abdel-Aal, *Computer Methods and Programs in Biomedicine* **1998**, *58*, 69-8.

23 A. Smellie, *Journal of Chemical Information and Modeling* **2004**, *44*, 1929-1935.