ASSOCIATION RULES OF DCI PATIENT CLUSTERS AND RELIABILITY OF CLUSTERING ANALYSIS

Baris Aksoy[1], Vincent Labatut[1], Murat Egi[1,4], Tamer Ozyigit[1], Petar Denoble[2], Costantino Balestra[3], Richard Vann[2], Alessandro Marroni[4]
[1]Galatasaray University Engineering & Technology Faculty, [2]The Duke Center for Hyperbaric Medicine and Environmental Physiology, [3]Environmental & Occupational Physiology Laboratory-Haute Ecole Paul Henri Spaak, [4]Divers Alert Network Europe Research Division

Introduction

Decompression illness (DCI) is an adverse outcome of decompression and has a wide spectrum of signs and symptoms. The treatment plan often depends on the classification of DCI, which makes the correct classification of DCI crucial; however, there is no consensus on the classification of DCI. [1-4]. We have previously attempted DCI clustering with statistical methods [5, 6] and suggested that data mining techniques can be used as a decision support tool to determine the type of DCI.

Our recent study was based on a classification of DCI using multivariate statistics to assess naturally associated clusters of signs and symptoms based on 1929 cases reported by hyperbaric chambers to the Divers Alert Network from 1999 to 2003 [7]. The aim of this study is to validate the reliability of the previous work by applying three different alternative clustering methods, by comparing the results of two-step clustering analysis with the Perceived Severity Index (PSI) [4] and to validate the characteristics of patient clusters using association rules. Additionally, we will present the most interesting association rules detected by the *A Priori* algorithm on the same data.

Methods

Below are the details of the algorithms we used as an alternative to our previous work:.

K-Means Algorithm

The k-means method is a widely used geometric clustering algorithm based on the article proposed by Lloyd in 1982 [7]. Given a set of *n* instances, the algorithm uses a local search approach to partition the instances into *k* clusters. A set of *k* initial cluster centers is chosen arbitrarily. Each instance is then assigned to the cluster whose center is the closest, and the centers are recomputed as centers of mass of their assigned points. This is repeated until the clusters stabilize. It can be shown that no partition occurs twice during the course of the algorithm, and so the algorithm is guaranteed to terminate. The k-means method is still very popular today, and it has been applied in a wide variety of areas ranging from computational biology to computer graphics [8]. One of the shortcomings of the k-means algorithm is the necessity to specify the number of clusters.

COBWEB Algorithm

Unlike the k-means algorithm which iterates over the whole dataset, the COBWEB Algorithm [9] incrementally incorporates instances into a classification tree. The incorporation of an instance is a process of classifying the object by descending the tree along an appropriate path, updating counts along the way, and performing one of several operators at each level. These operators include: classifying the instance with respect to an existing class, creating a new class, combining two classes into a single class (merging), and dividing a class into several classes (splitting). There are two parameters: *acuity* (minimum value to avoid infinite values) and *cutoff* (a parameter to suppress growth to avoid overwhelmingly large hierarchy) [10].

Expectation-maximization Algorithm

The Expectation-Maximization (EM) algorithm [11] is a statistical approach that makes use of the finite Gaussian mixtures model. A mixture is a set of *N* Gaussian probability

distributions, where each distribution represents a cluster. Additionally, a latent random variable models the probability for an instance to belong to a cluster. The algorithm has to estimate the parameters defining each cluster (i.e. mean and standard-deviation) and the latent variable distribution. The estimation process is iterative, with an expectation step consisting in using the current estimate of the parameters to process the expected value for the likelihood function, and a maximization step consisting in processing new estimates for the parameters by maximizing the previously processed likelihood. The algorithm terminates when cluster quality, i.e. expected likelihood, no longer shows significant improvement. The number of clusters can be either set by the user or estimated by the algorithm.

A Priori Algorithm

Agrawal et al. [7] introduced association rules by presenting an algorithm which generates significant logical rules involving items in a large database. An item is defined by a field and its value, for instance *X=x*. An association rule takes the form "if *X=x* then *Y=y*" where *X=x* and *Y=y* are items or item sets. Even for a small data set, many different association rules can be derived, so interest is restricted to those applying to a reasonably large number of instances, and having a reasonably high accuracy on the instances they apply to. The *support*, or coverage, of an association rule is the proportion of instances containing all the items used in the rule. Its *confidence*, or accuracy, is the ratio of the rule support to the support of its conditions (items located in the left part) alone. In other terms, the confidence corresponds to the proportion of instances for which the rule is correct, relatively to the total number of instances it applies to [10]. The association rule mining task is usually decomposed in two subtasks. One is to find item sets whose occurrences exceed a predefined threshold (minimum support) in the database, called frequent item sets. The second task is to generate association rules from those large item sets with a constraint of minimal confidence. The process starts with small frequent item sets, and then iteratively enlarge them until no frequent item set can be defined [12].

Data and Tools

DCI data were retrieved from the Diving Injury Database maintained by DAN America and based on reports provided by participating hyperbaric facilities using a standardized Diving Incident Report Form (DIRF). We used 1929 DIRFs (1368 males – 561 females) which were collected between 1999 and 2003. The average age of the patients is 37.94 with a range of 13 to 73. Twenty-five different signs and symptoms were listed: altered consciousness, confusion, dyspnea/chockes, cardiovascular signs (CV), pain, skin rash and mottling, localized skin swelling, abnormal sensations, hearing loss, vision problems, discoordination, muscular weakness, muscular problems, decreased skin sensitivity, bladder-bowel problems, headache, fatigue, nausea, dizziness, vertigo, paresthesia, tinnitus, numbness, paralysis and malaise. After a first exploratory analysis, an additional pain-only field was added, to improve performance for some algorithms.

The obtained clusters differ by their characteristic signs and symptoms. Our criterion was to look for clusters where each cluster have characteristic signs and symptoms that are distinct from other clusters. We favored those with features relevant with medical references. As a sample design, we used k-fold cross validation with 10 folds, which is the number of folds supposed to give best results in general [13]. We used Weka 3.5.7 [10], which is an open source software issued under the GNU GPL, as a data mining tool. We also used SPSS 16.0.1 for calculating the associations among the classifications and data transformation.

Results

K-Means

We applied the K-means algorithm for 2 to 6 clusters, and found out 4 clusters was the best parameter value according to our criteria. The estimated clusters contain 708, 637, 300 and 204 patients, respectively. Table 1 shows the signs and symptoms distribution over the

clusters. Numbness, paresthesia and decreased skin sensitivity are the most characteristic signs and symptoms for Cluster K1. Cluster K2 is characterized by fatigue, headache, skin, confusion, nausea, dyspnea/chokes, altered consciousness, vertigo, localized skin swelling, hearing loss and vision problems. Cluster K3 is the pain only cluster. Malaise, muscular weakness, paralysis and bladder bowel problems are the characteristics for Cluster K4.

Expectation-maximization

There is no need to specify the number of clusters for EM. The algorithm yielded four clusters containing 905, 253, 471 and 300 patients, respectively. We have set the maximum number of iterations to 100 but the iterations already repeated until convergence which did not exceed the maximum number of iterations. Table 1 shows the detailed results. Cluster E1 has numbness, paresthesia and decreased skin sensitivity as characteristic signs and symptoms. Weakness, muscular weakness, paralysis and bladder bowel problems are the characteristic signs and symptoms of Cluster E2. Cluster E3 is characterized by altered consciousness, confusion, dyspnea/chokes, localized skin swelling, skin rash and mottling, vertigo and hearing loss while Cluster E4 is the pain only cluster.

Signs and symptoms	K-means				EM				COBWEB		
	K1	K2	K3	K4	E1	E2	E3	E4	C1	C2	C3
Altered consciousness	5	**28**	0	10	6	10	**27**	0	11	8	**20**
Confusion	32	**78**	0	32	35	36	**71**	0	35	28	**77**
Dyspnea/chokes	27	**45**	0	20	19	26	**47**	0	27	10	**54**
CV	2	1	0	1	1	2	1	0	2	1	1
Pain	431	159	**300**	45	370	102	163	**300**	96	179	**652**
Skin Rash and Mottling	20	**89**	0	13	27	15	**80**	0	15	30	**76**
Localized Skin Swelling	4	**17**	0	1	4	1	**17**	0	1	5	**16**
Abnormal Sensations	12	17	0	8	10	10	17	0	10	10	17
Hearing Loss	2	**13**	0	1	1	2	**13**	0	1	4	**11**
Vision Problems	16	**26**	0	11	12	16	25	0	15	13	25
Discoordination	19	24	0	17	18	21	21	0	21	18	21
Muscular Weakness	63	7	2	**186**	9	**243**	4	2	**237**	10	11
Muscular Problems	22	25	0	5	23	9	20	0	7	19	26
Decreased Skin Sensitivity	**140**	48	0	60	**152**	72	24	0	73	95	76
Bladder Bowel Problems	7	2	0	**27**	1	**34**	1	0	**33**	2	1
Headache	40	**96**	0	21	59	22	76	0	24	49	**81**
Fatigue	59	**110**	0	26	77	29	89	0	32	60	**100**
Nausea	29	**57**	0	10	41	11	44	0	12	35	46
Dizziness	46	95	0	27	58	31	79	0	31	55	78
Vertigo	11	**28**	0	8	9	10	**28**	0	9	13	**25**
Paresthesia	**615**	51	0	74	**593**	132	15	0	128	**565**	26
Tinnitus	1	3	0	0	1	0	3	0	0	1	2
Numbness	**753**	135	0	115	**846**	148	9	0	163	**649**	167
Paralysis	44	8	0	**71**	8	**111**	4	0	**107**	5	11
Malaise	74	28	0	**204**	43	**237**	26	0	**255**	20	31

Table 1. Signs and symptoms for K-Means, EM and COBWEB clusters

COBWEB

We applied the COBWEB algorithm with acuity and cutoff parameters ranging from 0.1 to 1 and from 0.2 to 0.35, respectively. The best results according to our criteria were obtained for an acuity of 1 and a cutoff of 0.33. Three clusters were estimated, with 264, 721 and 944 patients respectively. Table 1 shows the characteristic signs and symptoms for each cluster.

Weakness, paralysis, muscular weakness and bladder bowel problems are the characteristic signs and symptoms for Cluster C1. Numbness and paresthesia are the signs and symptoms for Cluster C2. While pain is a significant sign for Cluster C3, it also has other signs and symptoms: confusion, altered consciousness, dyspnea/chokes, skin rash and mottling, localized skin swelling, hearing loss, headache, fatigue, vertigo. Compared to the other classification methods results, we did not find a pain-only cluster, which may be due to the incremental nature of the algorithm.

Analysis

We analyzed our results by carrying out both whole classification and cluster-to-cluster comparisons. First, we used Goodman-Kruskal's lambda, which measures the association level between nominal random variables, to compare classifications from the tested algorithms and PSI diagnosis, as shown in Table 2. The association between the classifications from the different algorithms is generally very strong, varying between 0.343 and 0.827. The association level is higher between the algorithms (including Two-step) than between the algorithms and PSI (manual classification). This means firstly that all automatic classification methods led to very similar results, and secondly that the resulting clusters are close to those defined by medical experts. COBWEB is the most different, because it only found three clusters. Two-step clustering is the most similar to PSI classification.

	EM	**COBWEB**	**K-Means**	**Two-Step**	**PSI**
EM	-	0.621	0.759	0.827	0.405
COBWEB	0.621	-	0.449	0.549	0.350
K-Means	0.759	0.449	-	0.802	0.343
Two-Step	0.827	0.549	0.802	-	0.444
PSI	0.405	0.350	0.343	0.444	-

Table 2. Association between clusters measured with Goodman-Kruskal's lambda

We performed a cluster-to-cluster comparison to check if these similarities were also present at this level. Table 3 shows the Pearson product-moment correlations between clusters from EM (E), k-means (K), COBWEB (C), Two-step (T), and PSI (P) classifications. Just by observing the characteristic signs and symptoms in Table 1, one may infer that correspondences exist between some clusters. This is confirmed by the strong correlations shown in Table 3, leading us to define four typical clusters. Typical cluster α correspond to clusters K1, E1, C2, T2 and P3 which are highly correlated (>0.9687), and all characterized by numbness and paresthesia. Typical cluster β corresponds to clusters K4, E2, C1, T3 and P1 which are highly correlated (>0.8808) and present muscular weakness, bladder bowel problems, paralysis and weakness. Typical cluster γ corresponds to clusters K2, E3 and T4, highly correlated (0.8120) and sharing altered consciousness, confusion, dyspnea/chokes, skin rash and mottling, localized skin swelling,, hearing loss and vertigo. Typical cluster δ is the pain only cluster, corresponds to clusters K3, E4, T1 and P4 (>0.9913). Cluster C3 is correlated to all K2,E3,T4 (>0,7441) and K3,E4,T1,P4 (0.9518), meaning it is likely a merge of typical clusters γ and δ. Cluster P2 is correlated with K2 (0.6230) and T4 (0.6065), and corresponds certainly to typical cluster γ, although it is not highly correlated with E3 (0.4928). Note that some PSI clusters are also non-trivially correlated to other clusters (E1:P1=0.6991, K1:P1=0.7387, E3:P4=0.7458), so the correspondence between PSI and the clustering methods is not straightforward.

Cluster α	**Cluster β**	**Cluster γ**	**Cluster δ**
E1:K1=0.9910	E2:K4=0.9814	E3:K2=0.8120	E4:K3=1
E1:C2=0.9817	E2:C1=0.9975	E3:C3=0,7864	E4:C3=0.9518
E1:T2=0.9957	E2:T3=0.9916	E3:T4=0.8693	E4:T1=1
E1:P3=0.9951	E2:P1=0.9151	E3:P2=0.4928	E4:P4=0.9913
K1:C2=0.9687	K4:C1=0.9881	K2:C3=0.7489	K3:C3=0.9518
K1:T2=0.9937	K4:T3=0.9738	K2:T4=0.9765	K3:T1=1

K1:P3=0.9933	K4:P1=0.8808	K2:P2=0.6230	K3:P4=0.9913
C2:T2=0.9815	C1:T3=0.9954	C3:P2=0.5802	C3:P4=0.9704
C2:P3=0.9730	C1:P1=0.9235	C3:T4=0.7441	C3:T1=0.9518
T2:P3=0.9950	T3:P1=0.9455	T4:P2=0.6065	T1:P4=0.9913

Table 3. Correlations between clusters measured with Pearson's correlation

A priori algorithm

We used the *A priori* algorithm to find out association rules linking symptoms and signs to the estimated cluster, in order to have a better understanding of the characteristic signs and symptoms. We applied the algorithm with minimum support and minimum confidence ranging from 0.1 to 0.9 and from 0.4 to 0.9, respectively. Our goal was to get simple rules, appropriate for human interpretation, and not necessarily the most accurate ones.

	Signs and symptoms	**Cluster**	**Cvrg**	**Cnf**
K-means	¬Altered consciousness, ¬Localized skin swelling, ¬Hearing loss, Numbness, ¬Malaise	K1	0.41	0.84
	¬Hearing loss, ¬Bladder Bowel problems, ¬Tinnitus, Numbness, ¬Malaise	K1	0.42	0.84
EM	¬Altered consciousness, ¬Localized skin swelling, ¬Hearing loss, ¬Muscular Weakness, ¬Vertigo, ¬Paralysis	E1	0.80	0.56
	¬Skin rash and mottling,¬Hearing loss, ¬ Muscular Weakness	E1	0.80	0.56
COBWEB	¬Altered consciousness, ¬Localized skin swelling, ¬Dyspnea/chokes,¬Muscular Weakness, ¬Bladder Bowel problems, ¬Vertigo	C2	0.79	0.45
	¬Malaise, ¬Paralysis, ¬Bladder Bowel problems, ¬Dyspnea/chokes, ¬Localized skin swelling	C2	0.58	0.45
Two-step	¬Consciousness, ¬Pulmonary, ¬Hearing, ¬Muscular Weakness, Numbness, ¬Weakness	T2	0.40	0.87
	¬Dyspnea/chokes, ¬CV, ¬Hearing loss, ¬Tinnitus, ¬Malaise, ¬Muscular Weakness, Numbness, ¬Bladder Bowel problems	T2	0.40	0.87

Table 4. Some of the association rules mined with the *A priori* algorithm

Table 4 shows the main rules mined for typical cluster α, for each algorithm where ¬ signs FALSE (does not exist). Interestingly, rather than the presence of cluster α most prominent features, *A priori* used the absence of the other clusters' features to discriminate the instances. Numbness is required for K-means and Two-step, but decreased skin sensitivity and paresthesia are not considered. Instead, instances must not present malaise (a cluster β feature) or hearing loss, altered consciousness, dyspnea/chokes, skin rash and mottling, localized skin swelling (γ cluster)., The rules mined for the other clusters are also in accordance with our interpretation of the clustering results: they focus on the characteristic symptoms and signs we found to be characteristic, either by requiring their presence or by requiring their absence in the case of those characteristic of another cluster (details documented elsewhere [14])

Discussion

In this work, we applied three clustering algorithms to the DCI classification problem. Each one is based on different approaches: K-means is distance-based, Cobweb is categorical and EM is probabilistic. We compared the clusters estimated by these three algorithms, those manually defined by the PSI medical experts [3] and those previously estimated by Ozyigit et al. by Two-step clustering [6]. We used Goodman-Kruskal's lambda to measure the association level between the different approaches, revealing all classifications shared similarities. We consequently performed a detailed cluster-to-cluster comparison with

Pearson's correlation coefficient. It appears some typical clusters characterized in terms of signs and symptoms appear whatever the considered classification method is. Typical cluster α is characterized by numbness and paresthesia. Typical cluster β is characterized by muscular weakness, bladder bowel problems, paralysis and malaise. Typical cluster γ is characterized by altered consciousness, confusion, dyspnea/chokes, skin rash and mottling, localized skin swelling, hearing loss and vertigo. Typical cluster δ is the pain only cluster, except for COBWEB, where this cluster was merged with typical cluster γ. To confirm these interpretations, we performed an association rule mining with the *A priori* algorithm, and focused only on rules with cluster as a conclusion. The results are consistent with our analysis, except the rules focus more on the absence of the signs and symptoms characteristic of other clusters, than on all the signs and symptoms we marked as characteristic.

The similarities observed between the estimated clusters and those defined manually show that different clustering methods can successfully classify DCI statistically. We can therefore conclude cluster analysis is a suitable technique to diagnose DCI type according to the patient's signs and symptoms. The long-term goal of this project is to optimize the treatment plan for a given DCI case, so the next step is to determine whether clustering is enough to define a treatment plan. Yet, we processed lambda between outcome and classifications and observed a close to zero association level (Documented elsewhere,[14]). This means diagnosis alone (be it manual or automatic) is not enough to predict the success or failure of the DCI treatment plan. Further studies will have to consider other data, such as patient age and sex, and treatment modalities to be able to perform a correct prediction. Such a prediction would then allow studying the relationships between symptoms, signs and other factors, and the outcome of the treatment plan, allowing to determination which data are relevant to select an adapted treatment plan.

References

1. Benton, M. and M. Glover, Dive Medicine. Travel Med Infect Dis, 2006. 4: p. 238-254.
2. Buch, A., et al., Cigarette Smoking and Decompression Illness Severity: A Retrospective Study in Recreational Divers. Aviat Space EnvironMed 2003. 74: p. 1271-1278.
3. Vann, R., et al., DAN's Annual Review on Decompression Illness, Diving Fatalities, and Project Dive Exploration, 2002.
4. Golding, F., et al., Decompression sickness during construction of the Dartford Tunnel. . Brit. J. Indus. Med., 1960. 17: p. 167-180.
5. Özyigit, T., et al., Empirical Classification of DCS patients using Cluster Analysis,Proceedings of the 33rd International Meeting of European Underwater and Baromedical Society 2007: Sharm el-Sheikh, Egypt. p. 213-217
6. Ozyigit, T., et al., Classification of DCI by Cluster Analysis. Proceedings of Undersea Hyperbaric Medical Society Annual Scientific Meeting, 2009, Las Vegas, P11.
7. Lloyd., S.P., Least Squares Quantization in PCM. IEEE Transactions on Information Theory, 1982. 28(2): p. 129-137.
8. Chiu , T., et al. A Robust and Scalable Clustering Algorithm for Mixed Type of Attributes in Large Database Environment. in Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2001.
9. Douglas, H.F., Knowledge acquisition via incremental conceptual clustering, Machine Learning Volume2, 1987. 2: p. 139-172.
10. Witten, I.H. and E. Frank, Data Mining: Practical machine learning tools and techniques with Java implementations. 2000, San Francisco, CA, USA,Morgan Kaufman
11. Kotsiantis, S. and D. Kanellopoulos. Gamma Ray Burst Search. Available from: http://grb.mnsu.edu/grbts/doc/manual/Expectation_Maximization_EM.htm.2009.
12. Kotsiantis, S. and D. Kanellopoulos, Association Rules Mining: A Recent Overview. GESTS International Transactions on Computer Science and Engineering, 2006. 32(1): p. 71-82.
13. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection in Proceedings of the International Joint Conference on Artificial Intelligence 1995.
14. Aksoy B. MSc Thesis. Computer Eng Masters Program, Galatasaray University, 2009.