Using Cluster Ensemble and Validation to Identify Subtypes of Pervasive Developmental Disorders

Jess J. Shen, BSc[1], Phil Hyoun Lee, MSc,[1], Jeanette J.A. Holden, PhD[2] and Hagit Shatkay, PhD[1]
[1]Computational Biology and Machine Learning Lab,
School of Computing, Queen's University, Kingston, Ontario
[2]Depts. of Psychiatry & Physiology, Queen's University, and Autism Research Program
Ongwanada Resource Centre, Kingston, Ontario

Abstract

Pervasive Developmental Disorders (PDD) *are neurodevelopmental disorders characterized by impairments in social interaction, communication and behavior.[1] Given the diversity and varying severity of PDD, diagnostic tools attempt to identify homogeneous subtypes within PDD. Identifying subtypes can lead to targeted etiology studies and to effective type-specific intervention. Cluster analysis can suggest coherent subsets in data; however, different methods and assumptions lead to different results. Several previous studies applied clustering to PDD data, varying in number and characteristics of the produced subtypes[19]. Most studies used a relatively small dataset (fewer than 150 subjects), and all applied only a single clustering method. Here we study a relatively large dataset (358 PDD patients), using an ensemble of three clustering methods. The results are evaluated using several validation methods, and consolidated through an integration step. Four clusters are identified, analyzed and compared to subtypes previously defined by the widely used diagnostic tool DSM-IV.[2]*

Introduction

Pervasive Developmental Disorders (*PDD*), also known as *Autism Spectrum Disorders*, are a group of neurodevelopmental disorders of varying severity affecting communication skills, social interaction, and behavior patterns.[1] Given the diversity of these conditions, current diagnostic tools, such as the widely used *DSM-IV* (*Diagnostic and Statistical Manual of Mental Disorders - 4th Edition*)[2], attempt to provide diagnostic criteria to divide PDD into relatively homogeneous subtypes by evaluating the three core areas it affects. The DSM-IV distinguishes among five categories: Childhood disintegrative disorder, Rett's disorder, Autistic disorder (autism), Pervasive Developmental Disorder – Not Otherwise Specified (PDD-NOS), and Asperger's disorder. The latter three are more common, while the first two are relatively rare.

The DSM-IV assigns one of the above conditions to patients, depending on whether a *cut-off threshold* for certain criteria is met or not. This *categorical* threshold-based partition suffers several shortcomings, including the use of an arbitrary threshold to distinguish normal from abnormal values, and the potential loss of important information through the use of such thresholds.[3,4]

Subtyping methods such as *Cluster Analysis*, which partition a dataset into subsets sharing common patterns, can evaluate an individual on a continuous scale of severity, with no categorical cut-off value designating a threshold between normal and abnormal.[4] Thus cluster analysis was previously suggested as an alternative for the dichotomous diagnostic criteria in DSM-IV.[5,19] Previous studies employing this idea have either used a very small dataset (30-50 cases)[6], used non-standard diagnostic tools making the results hard to apply[7], or included in their studies non-PDD patients with other developmental disorders.[8] Moreover, all these studies used a single clustering method, typically chosen ad-hoc, and most did not employ objective validation or evaluation methods.

When applying clustering methods to data, the discovered subsets can be arbitrary, and their meaning is not necessarily clear or useful. Typically clustering methods try to produce clusters that are compact (that is, items within clusters should be similar to each other), and well-separated (items in different clusters are expected to be significantly different from each other).[9] The area of *cluster validation* is concerned with evaluating clusters in a quantitative and objective way.[10] Formal methods examine how well a clustering fits a dataset (*fitness*) and how robust it is to perturbation in the data (*stability*).[11]

However, even these criteria do not guarantee that the results will be meaningful for the application domain, and different algorithms may lead to different partitions of the data into clusters. Recent studies on cluster validation attempt to circumvent the problem using cluster ensemble, which combines multiple clustering results into a single consensus solution.[9] This approach can improve clustering performance by consolidating the outputs on which several algorithms agree, typically leading to more robust results than those produced by any single method.[12] Moreover, the combination of multiple methods is more likely to expose the actual, domain-specific, trends present in the dataset.

In this study we analyze data from 358 PDD patients (referred to as *subjects*). For each patient the *Autism Diagnostic Interview-Revised* (*ADI-R*)[13] form, consisting of 93 questions, was filled and scored. It was then preprocessed to obtain 22 features per patient, and clustered using three different widely-used clustering

methods, as described below. The results were evaluated using fitness and stability criteria, and the 6 best clustering results form a *cluster ensemble*, consolidated through consensus clustering. The clusters are analyzed using statistical methods and domain knowledge, suggesting 4 stable subgroups that roughly correspond to – and further refine – 3 types commonly observed according to the DSM-IV diagnostic criteria.

Data and Methods

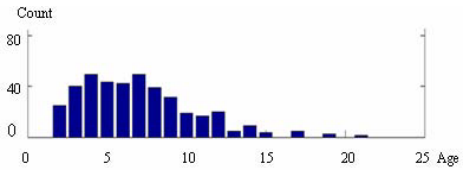
The data used in this study were collected from 394 PDD patients screened by the Autism Genetic Resource Exchange (AGRE). For each patient the *ADI-R* form, consisting of 93-questions, was filled and scored. This data was preprocessed to obtain 22 features per patient, and clustered using three widely-used methods, namely, *k*-means, *expectation-maximization*-based (*EM*) and *agglomerative hierarchical* clustering. The clusters obtained are evaluated using cluster validation methods, and the 6 best clustering results are integrated into a single solution through consensus clustering.

Data preprocessing: For each of the original 394 pre-screened PDD patients, 93 questions were answered and scored according to the ADI-R scoring algorithm[13]. Subjects with missing scores are excluded from this study, leaving *358* subjects whose ages range from 2 to 21 (mean = 6.9, std = 3.5, distribution shown in Fig. 1). Male to female ratio is 6:1. The ADI-R algorithm considers 39 of the questions to decide the patient's subgroup. Here we use much more of the data, specifically 64 questions are grouped into *22* non-redundant features, summarizing the information along certain typical dimensions, as shown in Table 1.[13,14] The questions that ask about relapse and age-specific manifestation of symptoms (and as such are only scored for the relevant age-group) are excluded from the original 93 questions, along with a few questions whose answers are highly correlated with each other. The feature values were obtained by summing the scores of the individual questions. The scores for all features were then normalized to the range 0-1, by mapping the maximum score to 1, the minimum to 0, and the intermediate values linearly to the (0,1) interval. To the best of our knowledge, this is the largest study done so far on subtyping PDD based on ADI-R data.

Clustering methods: We apply and compare three typical representative methods, widely used in practice:

- *k-means*[15,26]: Iteratively partitions the data into *k* clusters. In each iteration the centroid (typically, the mean) of each cluster is calculated, and each data point is assigned to the cluster with the closest centroid.
- *Hierarchical clustering*[16,27]: Starts by assigning each data point to a separate singleton cluster. Iteratively merges the closest pair of clusters. Merging is repeated until a pre-specified number of clusters, *k,* is obtained.
- *EM clustering*[17,18]: In contrast to the other two methods, EM is a probabilistic algorithm, assigning to each data point a probability to belong to each of the *k* clusters. The model used here is a set of *k* Gaussian distributions. Each iteration recalculates the Gaussian mean and variance, based on the probability of each point to belong to the cluster, locally maximizing the likelihood of the data given the cluster model.

Figure 1: Age distribution of the subjects



For all three methods, the value *k* ranged in our experiments between 3 and 7, as these are the number of clusters that have been used in previous studies.[19]

Table 1: Features Extracted from ADI-R (Version W-382D1)

Feature Name	Total Number of Questions (Questions shown in parenthesis)
1. Onset of symptoms	4 *(2, 4 ,86, 87)*
2. Early development	4 *(5, 6, 7, 8)*
3. Acquisition age of language	2 *(9,10)*
4. Conversational interchange	2 *(34, 35)*
5. Stereotyped speech	4 *(33, 36, 37, 38)*
6. Receptive communication	1 *(29)*
7. Gesture communication	4 *(42, 43, 44, 45)*
8. Behaviors to regulate interaction	3 *(50, 51 ,57)*
9. Peer relationships	3 *(49,62,63)*
10. Shared enjoyment	3 *(52, 53, 54)*
11. Socioemotional reciprocity	5 *(31 ,55 ,56,58,59)*
12. Social development	4 *(46, 47, 48, 61)*
13. Initiation of activities	1 *(60)*
14. Encompassing preoccupation	2 *(67, 76)*
15. Stereotyped motor mannerisms	2 *(77, 78)*
16. Ritualistic behavior	2 *(39, 70)*
17. Sensory issues	4 *(69, 71, 72, 73)*
18. Adherence to routine	2 *(74, 75)*
19. Symptoms of Rett's syndrome	2 *(79, 84)*
20. Aggression	3 *(81, 82, 83)*
21. Epilepsy	1 *(85)*
22. Demonstrated isolated skills	6 *(88, 89, 90, 91, 92, 93)*

Cluster validation: As we compare results obtained while varying the number of clusters, *k*, *(3≤k≤7)*, the results were evaluated based on two measures: *fitness* and *stability*, mentioned earlier. Fitness for *k*-means and hierarchical clustering is measured using the *Mean Silhouette Width*[20], which is the ratio between the average distance among items within the same cluster (compactness) and the distance among items not in the same cluster (separation). A high mean silhouette width is desirable, as it indicates tight clusters that are well-separated from each other. Formally, the mean silhouette width for a clustering with *k* clusters is denoted by S_{mean}, and defined as:

$$S_{mean} = \frac{\sum_{j=1}^{k} \sum_{i \in C_j} \frac{b_{ji} - a_{ji}}{\max(a_{ji}, b_{ji})}}{\sum_{j=1}^{k} |C_j|},$$

where a_{ji} is the average distance from the i^{th} point in cluster C_j to all other members of C_j, and b_{ji} is the average distance from the i^{th} point in cluster C_j to members of its closest neighboring cluster $C_l \neq C_j$.

For the EM method, as cluster assignment is probabilistic, we used the *Bayesian information criterion* (*BIC*) defined as:

$$\mathrm{BIC} = -2L + v \cdot \ln(n),$$

where n is the number of data points, L is the likelihood (the probability of the data given the clustering model), and v is the number of free parameters in the model. A model that has a lower BIC is preferred.[21]

Stability was evaluated based on replication analysis, as proposed by Breckenridge.[22] The dataset is split into two equal subsets denoted $C_{training}$ and C_{test}. Each of the subsets is partitioned into k clusters using any clustering method. The clusters obtained for the $C_{training}$ set are viewed as "ground truth", and supervised learning is used to train a classifier based on these clusters, where the cluster labels are viewed as the classes. In this work we use a standard *Random Forests* classifier, as it was shown to be highly accurate in a variety of cases.[23]

The classifier is then used to classify the C_{test} set,and the agreement between the cluster labels assigned to C_{test} data by the unsupervised clustering algorithm and by the Random-Forest classifier is calculated. The agreement is measured by the *adjusted Rand index* (*ARI*),[24] which is calculated as described next.

For a dataset X, let C and P be two partitions, $C=\{C_1,..C_N\}$ and $P=\{P_1,..P_M\}$, where

$$X = \bigcup_{i=1}^{N} C_i = \bigcup_{j=1}^{M} P_j .$$

Let $x_i \in X$ be a data point in X. We denote by $C(x_i)$ the subset C_k, to which x_i belongs under partition C and $P(x_i)$ the subset P_l, to which x_i belongs under partition P. Let A be the set of pairs of points $x_i, x_j \in X$ that are placed in the *same* subset according to both partitions, formally:

$A = \{\langle x_i, x_j\rangle \mid C(x_i) = C(x_j) \text{ AND } P(x_i) = P(x_j)\}$,

and D be the set of pairs of points $x_i, x_j \in X$, that are placed in *different* subsets according to both partitions, formally: $D = \{\langle x_i, x_j\rangle \mid C(x_i) \neq C(x_j) \text{ AND } P(x_i) \neq P(x_j)\}$.

Denote by $|A|$ and $|D|$ the number of pairs in the sets A and D, respectively. The adjusted Rand index (*ARI*) is defined as:

$$ARI = \frac{R - E(R)}{1 - E(R)},$$

where $R = (|A|+|D|)/(\#\text{ of pairs } \langle x_i, x_j\rangle \text{ in } X)$, $E(R)$ is the expected value for R under chance agreement between C and P, and 1 is the maximum value that R can obtain (when C and P the same). The value of the *ARI* ranges between -1 and 1. The larger the *ARI* is the better the agreement between the partitions C and P.

For each of the clustering methods, the replication analysis process was repeated using about 200 random splits of the data, and the number of clusters k for which the ARI values are statistically significantly[a] larger than those produced for any other k, was taken as the optimal number of clusters, denoted k_{opt}. The best clustering solution with k_{opt} clusters was used as a component in the cluster ensemble in this study.

Applying all three clustering methods and choosing the best solutions according to the two criteria (stability and fitness) gives rise to three pairs of best clustering solutions (the most stable and the most fit solutions, for each of the clustering methods). When the two criteria arrive at the same number of clusters k, the pair of solutions for each clustering method consists of two identical clustering solutions. We consider them both as distinct solutions in the integration step, as two separate criteria identified this solution as optimal, and as such it should carry twice the weight in the integration step.

Cluster Integration: To arrive at a unique clustering solution based on the 6 best solutions described above, each subject is represented as a 6-dimensional vector, where the i^{th} position in the vector is the cluster label assigned to the subject by the i^{th} clustering solution (where $1 \leq i \leq 6$). We call these vectors *prototype vectors*. The 6-dimensional vectors obtained are shown in Table 3, along with the number of subjects represented by each vector. The 6 values in each (column) vector have no numerical interpretation, as they denote cluster labels. Therefore k-modes clustering[25], a variation of the k-means applicable to non-numerical data, was used to produce a single stable consensus clustering solution. The metric used to evaluate the distance of each vector to the mode of its cluster is the Hamming distance, which essentially counts the number of positions on which two vectors disagree. Future studies will experiment with other consensus clustering methods and measures.

Results

Table 2 shows the optimal number of clusters, k_{opt}, obtained for each clustering method, based on the fitness and stability tests. About 200 random splits of the data were used for each of the clustering methods, thus obtaining statistically significant differences, ($p \leq 0.05$), between the *ARI* calculated for the different number of clusters k (where $3 \leq k \leq 7$).

While for each of the methods, the 3-cluster solution produced is considered to be optimal in almost all cases (the exception is the 5-cluster solution which is most stable for hierarchical clustering), the different 3-cluster solutions produced by the different algorithms typically do not agree with one another. This is a well-known problem of using a single clustering method in cluster-analysis, and justifies our use of an ensemble to resolve the differences. To obtain a unique and unified solution, the ensemble of clustering results is integrated through consensus clustering.

[a] Significance measured using the *Wilcoxon signed-rank test*.

Table 2: k_{opt} from Cluster Validation

	k-means	Hierarchical	EM
Fitness	3	3	3
Stability	3	5	3

Clustering Integration:

Table 3 shows the distribution of the 6-dimensional vectors across the subjects. The table shows that only 19 of all possible 6-dimensional combinations of cluster assignments were actually obtained and that the vast majority of the subjects (297) are represented by the 4 highly distinct prototype vectors shown as the 4 left columns of the table. The subject distribution based on their respective prototype vectors is thus clearly 4-modal.

Table 3: Frequencies of prototype (column) vectors. Each entry corresponds to the cluster assignment according to the most fit (*f*) and most stable (*s*) solutions based on each of the three clustering methods, *k-means, Hierarchical* and *EM.* The four modes are shown in boldface.

Prototype No.			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Proto. Vector	*k-means*	*f*	1	2	3	3	3	1	2	2	1	2	2	3	3	1	2	2	2	3	3
		s	1	2	3	3	3	1	2	2	1	2	2	3	3	1	2	2	2	3	3
	Hier.	*f*	1	1	1	1	1	1	3	1	1	1	2	1	2	1	1	1	1	1	2
		s	1	2	2	2	1	1	5	1	1	2	4	1	3	2	1	1	2	1	4
	EM	*f*	3	2	2	1	1	2	2	2	1	1	2	2	1	1	1	3	3	3	1
		s	3	2	2	1	1	2	2	2	1	1	2	2	1	1	1	3	3	3	1
Subject Freq.			**123**	**90**	**52**	**32**	12	9	8	6	5	5	4	3	3	1	1	1	1	1	1

We thus obtain a partition of the data into 4 main clusters, whose modes are the 4 leftmost prototype vectors in Table 3. Each of the remaining 61 subjects is assigned to the cluster whose mode is closest to the subject's corresponding prototype vector, based on the Hamming distance between prototypes. The exact same clustering is obtained by the 4-mode algorithm.

Figure 2 shows the feature values associated with each of the 4 clusters. The rows correspond to the 22 features (listed in Table 1), and the columns are the cluster labels. The number of subjects in each cluster is shown underneath each column. Feature values range from 0 to 1, and darker shades correspond to higher values for each of the features. The Wilcoxon rank-sum test is used to ensure that the difference in feature-value distribution is indeed statistically significantly different between every pair of clusters ($p \leq 0.05$).

Relation to clinical diagnoses: For 210 of the 358 subjects studied here, the Autism Genetic Resource Exchange (AGRE) provides clinical diagnoses that were made by a physician based on DSM-IV criteria. These diagnoses include: Autism, PDD, PDD-NOS (PDD-Not-Otherwise-Specified), Asperger's syndrome and a few cases of Attention Deficit Hyperactivity Disorder (ADHD).

Table 4 shows the correspondence between the clinical diagnoses and the cluster memberships. The highest number of subjects sharing a diagnosis within each cluster is shown in bold. Clusters 1, 3 and 4 are dominated by Autism diagnoses while Cluster 2 is dominated by Asperger's syndrome. The Chi-square test validates that the distributions of clinical diagnoses are statistically significantly different ($p \leq 0.05$) between clusters 1, 2 and either 3 or 4. (The diagnoses distributions for patients in clusters 3 and 4 are not statistically significantly different from one another).

Figure 2: Feature Profile for the *4*-Cluster Consensus Solution.

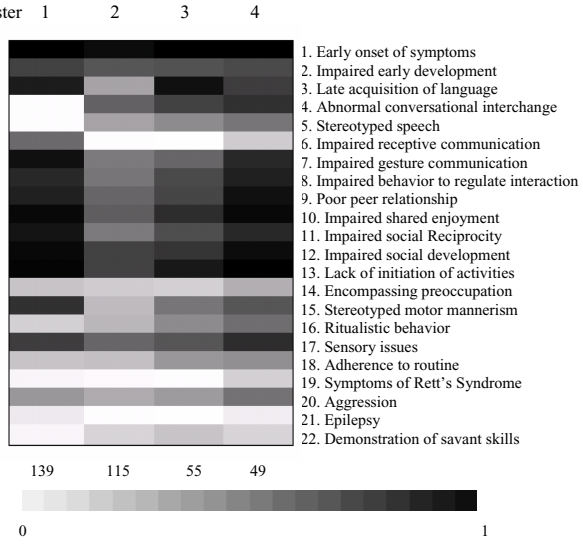


Table 4: Clinical Diagnoses and Clusters

	1	2	3	4	Total
Autism	**76**	19	**14**	**16**	125
Asperger's	0	**21**	2	4	27
PDD-NOS	2	9	2	3	16
PDD	10	17	7	6	40
ADHD	0	0	1	1	2
Total	88	66	26	30	210

Discussion

Each of the clusters shown in Fig. 2 is characterized by a distinct distribution of feature values, where Cluster 2 is clearly different from the other three, while the differences among Clusters 1, 3 and 4 are more subtle.

Cluster 2 (denoted C_2) contains the least impaired subjects, and, as shown in Table 4, includes most of the Asperger's syndrome patients – a relatively mild form of PDD. Notably, subjects in C_2, show a close-to-normal age of language acquisition (feature *3*, *f3* for short), and the onset of the abnormalities in this group is late. While most of the symptoms are less severe in this group, social abilities and communication (*f4* ,*f7-13*) are clearly impaired, and patients are hypersensitive to stimuli (*f17*).

Clusters C_1 and C_4 group the most severely impaired patients, corresponding to the typical *Autism* subtype in DSM-IV. Both clusters show highly impaired social functions (*f8-13*), and a higher tendency to epilepsy (*f21)* compared with C_2 and C_3. However, the two clusters are still distinct.

Subjects in C_1 demonstrate late language acquisition and language impairment so severe that they are considered non-verbal, have no functional use of three-

word phrases, and at times are completely mute. Notably, their score on features *4* and *5* is 0, as they cannot be evaluated due to the lack of speech. Almost 2/3 of them have problems understanding other people's language (*f6*), and their social reciprocity is significantly worse than that of C_4 subjects. Stereotyped motor mannerisms (*f15*) – possibly as a result of their very limited communication skills – and hypersensitivity to stimuli (*f17*) are apparent.

In contrast, the subjects in cluster C_4 are characterized by overly persistent (*f16,18*) as well as more aggressive behaviors (*f20*). While they do demonstrate some language skills, their verbal development is delayed (*f3*), and severely impaired as shown by stereotyped speech and poor conversational ability (*f4,5*).

Cluster C_3 is characterized by an intermediate level of severity. It is similar in characteristics to C_4, but shows lower scores for almost all features except for delayed language acquisition (*f3*). For 12 of the features, the lower scores compared to C_4 are highly statistically significant ($p \leq 0.05$). We note that the characteristics of the subjects in this cluster are generally those of PDD-NOS, and the fact that the cluster contains subjects that were not diagnosed as such, highlights the value of cluster-analysis as a method for identifying subtypes in the data that may not be identified using a rule-based algorithm such as that defined for ADI-R.

Conclusion

Using an ensemble of clustering methods and cluster validation, we identified four clusters that roughly correspond to – and further refine – three main subtypes of PDD, namely Autism, PDD-NOS and Asperger's syndrome. The dataset used here is the largest ADI-R dataset analyzed so far. We note that our clusters are characterized by a distribution of scores along many questions and features, and thus distinguish among subgroups based on finer criteria than those defined by DSM-IV. The clusters form a continuum of severity along the different impairments and thus agree with the opinion held by many researchers that PDD subtypes should not be distinguished based on discrete, mutually exclusive, impairments but rather form a spectrum of disorders varying in severity from almost normal to highly impaired.[19] In future studies we plan to examine other clustering consensus strategies, as well as integrate other forms of information (such as genomic data, IQ data, family information etc.) into the clustering process.

Acknowledgements

We are grateful to Heidi Penning, Dr. Ira Cohen, Dr. Xudong Lee and Melissa Hudson for their help and insight in interpreting and analyzing the ADI-R data. The work was supported by HS's NSERC Discovery grant # 298292-04 and CFI New Opportunities award #10437, and by JJAH's CIHR grant #43820 and OMHF grant.

References

1. Strock M. Autism spectrum disorders (Pervasive Developmental Disorders). NIH Pub. No. 04-5511, NIMH.
2. The American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders - 4th Ed. 1994.
3. Skodol A, Oldham J *et al.* Dimensional Representations of DSM-IV Personality Disorders: Relationships to Functional Impairment. AJP, 2005; 162:1919-25.
4. Kessler RC. The Categorical versus Dimensional Assessment Controversy in the Sociology of Mental Illness. J. of Health and Social Behavior. 2002.
5. Myhr G. Autism and other pervasive developmental disorders: exploring the dimensional view. Canadian Journal of Psychiatry. 1998; 43:589–95.
6. Stone W, Ousley O, Hepburn S, Hogan K, Brown C. Patterns of adaptive behavior in very young children with autism. AJ MR. 1999; 104: 187–199.
7. Siegel B, Anders T, Ciaranello R, Bienenstock B, Kraemer H. Empirically derived subclassification of the autistic syndrome. JADD. 1986; 16: 275-294.
8. Rescorla, L. Cluster analytic identification of autistic preschoolers. JADD. 1988; 18: 475–492.
9. Handl J, Knowles J. Multiobjective clustering and cluster validation. Springer Series on Computational Intelligence. 2006; 16: 21-47.
10. Jain AK, Dubes R. Algorithms for Clustering Data. Prentice-Hall. 1998.
11. Tan P, Steinbach M, Kumar V. Introduction to Data Mining. Pearson Addison Wesley. 2005.
12. Topchy A, Law M, Jain A, Fred A. Analysis of Consensus Partition in Cluster Ensemble. ICDM. 225-232. 2004.
13. Lord C, Rutter , Le Couteur A. ADI-R, Autism diagnostic interview-revised. J. Autism Dev Disord. 1994; 24:659-685.
14. Penning H. and Cohen I. Personal Communications. 2006.
15. MacQueen J. Some methods for classification and analysis of multivariate observations. Proc. of the 5th Symp. on Mathematical Statistics and Probability. 1967; 1: 281–297.
16. Johnson SC. Hierarchical clustering schemes. Psychometrika, 1967; 32: 241–254.
17. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. J. of the Royal Statistical Society, Series B. 1977; 39(1):1–38.
18. Witten I, Frank E. Data mining: practical machine learning tools and techniques, 2nd Ed. Morgan Kaufmann. 2005. (WEKA implementation of machine learning tools).
19. Beglinger LJ, Tristram HS. A review of subtyping in autism and proposed dimensional classification model. Journal of Autism and Developmental Disorders. 2001; 31(4):411-22.
20 Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis, Wiley. 1990.
21. Raftery A. A note on Bayes factors for log-linear contingency table models with vague prior information, J. of the Royal Statistical Society. 1986; 48(2): 249-250.
22. Breckenridge J. Replicating cluster analysis: Method, consistency, and validity. Multivariate Behavioral Research. 1989; 24: 147- 161.
23. Breiman L. Random Forests. Technical Report 567. Berkeley. Dept. of Statistics. U. of California. 1999.
24. Arabie, H. Comparing partitions. J. of classification, 1985; 2: 193-218
25. Huang Z. Extensions to the *k*-means algorithm for clustering large data sets with categorical values. Data Mining Knowledge Discovery. 1998; 2(2): 283–304.
26. Toknomo K. Matlab implementation of the k-means. *http://people.revoledu.com/kardi/tutorial/kMean/matlab_kMeans.htm.*
27. The MathWorks Inc. MATLAB version 7. R14. 2005.