p-ISSN : 2335-1357 e-ISSN : 2437-1122

Mediterranean Journal of Modeling and Simulation

Med. J. Model. Simul. 06 (2016) 001-011

Ensemble classification methods for autism disordered speech

Zoubir Abdeslem Benselama [a] *, Mohamed A. Bencherif [b] , Abderrezak Guessoum [a] , Mohamed A. Mekhtiche [b]

[a] Electronic Department, Saad Dahleb University, Blida(09000), Algeria.
[b] King Saud University, CS2R, Riyadh, Kingdom of Saudi Arabia.

ARTICLE INFO

Article history :
Received March 2016
Accepted May 2016

Keywords :
Autism ;
Pathology ;
Speech disorder ;
Feature selection ;
Ensemble classifiers.

ABSTRACT

In this paper, we present the results of our investigation on Autism classification by applying ensemble classifiers to disordered speech signals. The aim is to distinguish between Autism sub-classes by comparing an ensemble combining three decision methods, the sequential minimization optimization (SMO) algorithm, the random forests (RF), and the feature-subspace aggregating approach (Feating). The conducted experiments allowed a reduction of 30% of the feature space with an accuracy increase over the baseline of 8.66% in the development set and 6.62% in the test set.

1. Introduction

Autism is a term for a wide range of developmental brain disorders, called autism spectrum disorder (ASD) in the scientific community. The term spectrum refers to a collection of symptoms, skills, and levels of impairment or disability. Some individuals are impaired whilst others are severely disabled [1]. According to [2], ASD is sometimes called pervasive developmental disorder (PDD), and has been classified into five major classes :

1. Autistic disorder (classic autism)
2. Asperger's disorder (Asperger syndrome)
3. Pervasive developmental disorder not otherwise specified (PDD-NOS)
4. Rett's disorder (Rett syndrome)
5. Childhood disintegrative disorder (CDD).

The symptoms of ASD vary from one child to another, but can be classified into three areas :

*Email : benselamaabd@hotmail.com

1. Social impairment
2. Communication difficulties
3. Repetitive and stereotyped behaviors.

In general, parents are the first to notice the abnormal behavior of their child. Sometimes ASD can be found in very young babies, when the infant starts focusing on fixed objects and fails to engage in play with his or her parents. Sometimes children behave normally until the age of two or three, at which point the symptoms of ASD appear, such as being silent, unsocial, indifferent, and displaying a loss of development (which is called regression).

Aiming to contribute to the early detection of speech impairments, many hospitals and speech departments have recorded speech databases in order to automatize the process of pathology detection and classification. Likewise, many research papers have also dealt with the detection of impaired speech, such as [3] and [4] on stigmatism classification, [5] on prosodic assessment of language impaired children, and [6] on automatic classification.

The need to investigate using computerized automatic methods requires assessed recorded pathological databases. For this autism related work, the Child Pathological Speech Database (CPSD) has been used; this database was recorded in two university departments (pediatrics and psychiatry) in Paris, France. The first is located at the "Université de Pierre et Marie Curie/Pitié Salpêtrière Hospital, while the second department belongs to the Université Rene Descartes/Necker Hospital.

The database consists of 99 children aged from 7 to 19 years and of both genders. The pathological database has been segmented into two main classes, defined as typical and atypical autism, and a second deep segmentation includes PDD, Dysphasia (DYS), and Not-Otherwise Specified (NOS). The set of distributed recorded files is presented in Table 1. [7].

Table 1 – CPSD pathological speech database distribution.

	Autism	Train	Dev.	Test	Total
Typical (TYP)	TYP	566	543	542	1651
Atypical (ATY)	PDD	104	104	99	307
	NOS	104	68	75	247
	DYS	129	104	104	337
	Total	903	819	820	2542

In section 2, we will describe the classification methods, and section 3 presents the feature selection scheme. The implementations and results are described in section 4, which is followed by a discussion in section 5, before the paper concludes.

2. Classification methodology

Classifiers have the ability to split the space of features into low-level boundary spaces, thus allowing an expert decision of the probability of a feature vector belonging to one

or more subspaces. The error in the decision is more related to the correlation of the feature space and the overlap between the sub/spaces ; in this order of idea, the use of multiple experts improves the "point of view" of the decision and lowers the probability of belonging to more than one space. Many classifiers, such as support vector machines and decision trees, can show exceptional results on some datasets and very low accuracies on others. In this conjecture, using different voters can handle the disparity between the classifiers ; the better approach is to have an expert for each subspace or class. Unfortunately, with the increasing number of classes and problems, other decision methods have to be implemented.

In this paper we have opted for strong and weak classifiers, and experiments showed that by a tuned voting principle, the overall accuracy is better than each classifier alone.

2.1. Sequential minimization optimization (SMO) algorithm

The John Platt's SMO algorithm for training a support vector classifier has been investigated in the pathological or emotional context [7]. The support vector machines have tremendously shown their ability to use intrinsically high dimensional hyper-planes to separate classes using binary splits. In such situations, the problem is to find a solution to the optimization equation [8]

$$\min_{w,b,\xi} \left\{ \frac{1}{2} \|w\|^2 + C. \sum_i \xi_i \right\} \tag{1}$$

under the constraints defined by :

$$l_i \left(w.x_i - b \right) \geq 1 - \xi_i, \ 1 \leq i \leq n, \ \xi_i \geq 0 \tag{2}$$

where C is the penalty for mislabeled examples and n the number of training files within the dataset. Once the model is built, it can be generalized to the development and test sets.

In our experiments, a polynomial kernel of degree one was used, as shown in equation 3.

$$K \left(x, y \right) = \langle x, y \rangle \tag{3}$$

2.2. Random Forest (RF)

In [9], Breiman proposed a variant of bagging called random forests (RF), which is an ensemble of decision trees built upon independent and identically distributed random vectors induced in a growing decision tree. Each tree uses a set of m features selected from the whole set of features, and grows until convergence. The sub-trees use an ensemble technique to decide on the class of the new instance.

The RF model in [10] is a predictor of a set of regression trees $r_n\{X, \Theta_m, D_n, m \geq 1\}$, where X are the random variables, and the Θ_i, $i = 1...m$ are i.i.d. outputs issued from a randomized variable Θ. The set of trees are then aggregated or combined to form the regression estimation defined as :

$$\bar{r}_n(X, D_n) = E_\Theta [r_n(X, \Theta_m, D_n)] \quad (4)$$

where E_Θ is the expected value of the random parameter X and the data D_n.

Each individual random tree will be built in the following manner :

At each node, a coordinate of X, from the d dimension vector is selected, with the k-th feature having a probability $p_{n,k}$ of being selected. Once the coordinate is chosen, a division or split is initiated at the midpoint of the selected side.

The randomized tree $r_n(X, \Theta_m, D_n)$ generates the output for which the corresponding vector X_i falls within the same cluster of the random partition as X.

Each individual tree will contain approximately k_n terminal nodes and each single leaf will have a Lebesgue measure of $1/k_n$. If X has a uniform distribution on the interval $[0, 1]^d$, it will result in n/k_n observations per terminal node.

2.3. Feature subspace aggregating ('Feating')

The technique is itself an ensemble approach [11] ; it is a generic concept that can enhance the predictive performance of learners, and it is a generalized form of the Average One-Dependence Estimators (AODE) method. It uses a local model rather than a global one, and is formed by splitting the feature sub-space into non-overlapping local regions and ensuring that different subdivisions provide the distinct local neighborhoods for each point in the feature space. The problem is tackled by [11 in the way that solving a small aggregated problem is easier than solving a global problem.

The proposed feature-subspaces, issued from exhaustive subdivisions, are the backbone of an ensemble method that groups or aggregates all the sub-models known as local models, or a randomized part of them.

The feating is based on the following algorithm 1 :

Algorithm 1 : Feating (D, A, h)

– Build a set of Level Trees based on Feating

INPUT D : Training set, A : Set of given attributes, h : Maximum Level Tree

OUTPUT E : Collection of Level Trees

$E \longleftarrow$ Start by an empty tree, $n \longleftarrow |A|$ /* Number of features.

$N \longleftarrow C_n^h$, $P \longleftarrow$ rankAttribute(A),

for $i = 1$ to N do /* Construct an attribute list from P based on index i */

$L \longleftarrow$ attributeList(P, i),

$E \longleftarrow E \cup$ BuildLevelTree(D, L, 0),

end for

Return E

The feating technique has two main advantages :

– Decreases the execution time as the level of localization is increased

– Best fit for large data size as in our case.

2.4. Voting techniques

In decision theory, combining classifiers rests on two main schemes :

a- Use of optimal (sub-optimal) basic classifiers.

b- Use of strong combination rules.

Many classifiers pretend good accuracies over the training data, essentially due to large training data and repeated training sessions, leading unfortunately to a limited generalization process over newer test data. In order to avoid this type of problem, let us start with the assumption that each classifier uses a dissimilar approach to tackle the training data. The decision will be "classifier dependent" and tends to be more favorable to part of the data rather than to the other parts. Thus, adding different classifiers or expert decision makers will improve the decision, under the constraint of having strong combining rules : "The use of combination of multiple classifiers was demonstrated to be effective, under some conditions, for several pattern recognition applications" [12].

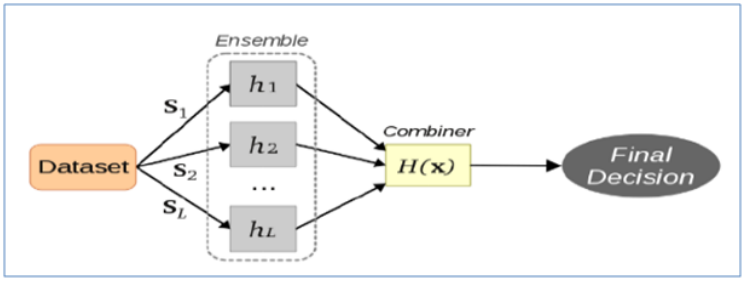

Fig. 1 – Ensemble selection in a parallel scheme.

Fixed rules such as majority vote, minimum, and maximum probability rules have been tested and show performance increase in the development set. The majority rule encompasses that classifiers can decide on an autism case in a majoritarian manner, and there are cases where the majority vote [12] can lead to a decrease of the overall accuracy ; the highest probability supposes that an expert per class can win. Other rules are also listed in the experiment tables, but they fluctuate between majority and maximum probabilities.

3. Features

The different speech features have been generated from the opensmile software [13]. The precompiled configurations included in the software contain different combinations of features, as in the proposed TUM baseline [7]. These features follow the Attribute Relation File Format (ARFF) and can be used in the Weka data mining java platform [14].

The predefined speech features are also called low-level descriptors (LLD), as they describe the basic features of speech such as the MFCC, the LPC, the ZCR, and the voice probability. All the LLD parameters are shown in Table 2.

Table 2 – Speech low-level descriptors (LLD).

LLD	**Process / parameters**	**Qty.**
Log Energy	After Hamming windowing and pre-emphasis (0.97)	1
MFCC 0-12	Pre-emphasis 0.97, Ham. window	13
Critical band spectrum	Over 26bands	26
Zero crossing Rate	Frames of 25ms,10ms overlap	1
Voice Probability		1
F0	F0+F0envelope	2
Spectral band energies	[0-250], [0-650],[250-650], [1000-4000][3010-9123]	5
Spectral	Roll-Off Point 25 , 50, 75,90	4
Spectral Flux	Over successive frames	1
Spectrum	Spectral Centroid, Max, Min, Energy	4
	Total	58

The LLD parameters are smoothed by a moving average filter of length three before being sent to a regression module, in order to compute the delta regression coefficients from the data contour. Then, statistical functional methods are applied, and the total number of coefficients is computed as follows : (58 LLD+58 DELTA_LLD)*39Functionals=4524, as presented in Table 3.

Table 3 – LLD Functionals.

Functionals	**Type**	**Qty.**
Extremes	Max position, min position, amplitude, norm per frame	5
Regression	Linear regression coefficients, centroid, quadratic error, quadratic regression	9
Moments	Variance, std. dev., skewness, kurtosis,	5
Percentiles	Quartiles, inter quartile, percentile (0.95, 0.98)	8
Crossings	Zero crossing rate	1
Peaks	Number of peaks, mean peak distance.	4
Means	Mean, abs. mean, non-zero mean, non-zero geometric mean	7
	Total	39

A detailed view of the spanning features is presented in Fig. 2., where the input wave file is fed to different blocks such as framing and vector emphasizing. Then, all the data are collected into a smoother and a regression module, and finally all types of functionals are generated and output to Weka.

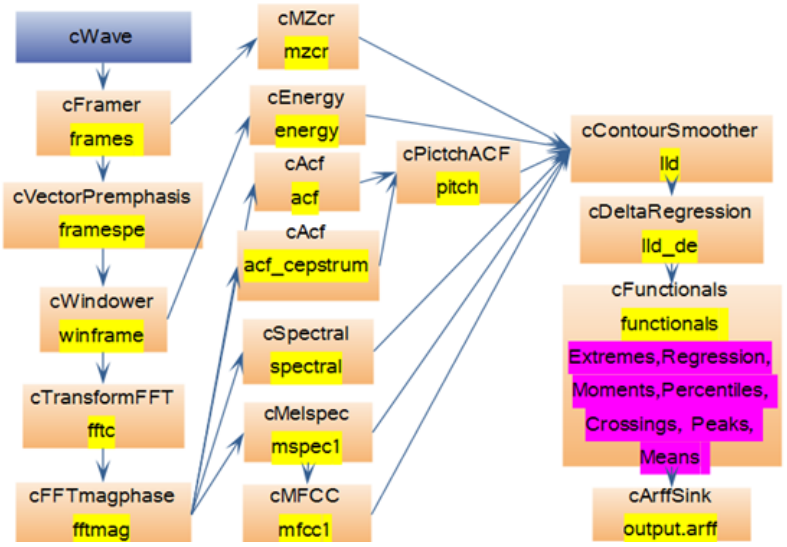

Fig. 2 – Selected features flowchart computation.

4. Experimentation

In all the following experiments, three datasets are used. The train and development datasets have known classes while the test set has unknown classes, and the TUM website generates the accuracy of the test set for each of our models.

As an initial baseline investigation, the SMO has been adopted, with penalty parameters ranging from 0.0001 to 0.15, with different opensmile speech configurations, as illustrated in Fig. 4.

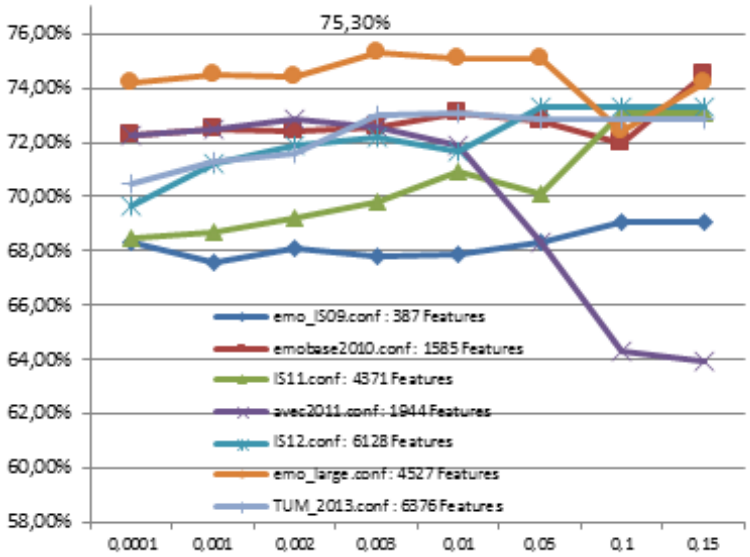

Fig. 3 – SMO recognition accuracies for the autism-diagnosis using different speech configurations.

The SMO best per-class results are shown in Table 4.

Table 4 – SMO development set results.

	DYS	NOS	PDD	TYP	Sum
DYS	71	6	16	11	104
NOS	12	16	17	23	68
PDD	25	30	34	15	84
TYP	3	14	30	496	543
Autism (Diagnosis) total accuracy = 75.34%					

The same SMO model has been applied to the test set, giving an accuracy of 75.61%, with an increase of 5.81% over the TUM baseline, as presented in Table 5.

Table 5 – Test set results using SMO.

	DYS	NOS	PDD	TYP	Sum
DYS	38	4	35	27	104
NOS	0	37	9	29	75
PDD	24	17	25	33	99
TYP	3	7	12	250	542
Autism (Diagnosis) total accuracy = 75.61%					

The different classifiers (SMO, RF, and Feating) have been trained and tested independently and then embedded in a vote module, as shown in Fig. 6. Let us remark that the classifiers have been added to the vote process incrementally. In order to see the effect of incremental vote process, the development results of the SMO-RF are presented in Table 6.

Table 6 – Classification scheme using the vote process on the development set (819 instances).

	Development set							
	Single classifiers		Ensemble voting classifiers					
	SMO	RF	Maj. Vote	Avg. Prob.	Maj. vote	Product Prob.	Min prob.	Max prob.
Correctly classified	629	600	612	632	611	630	619	641
Incorrectly classified	190	219	207	187	208	189	200	178
Kappa statistic	0.546	0.390	0.471	0.515	0.468	0.507	0.476	0.546
Mean absolute error	0.281	0.211	0.1264	0.2462	0.127	0.163	0.208	0.264
Accuracy (%)	76.80	73.26	74.72	**77.16**	74.60	76.92	75.58	**78.26**

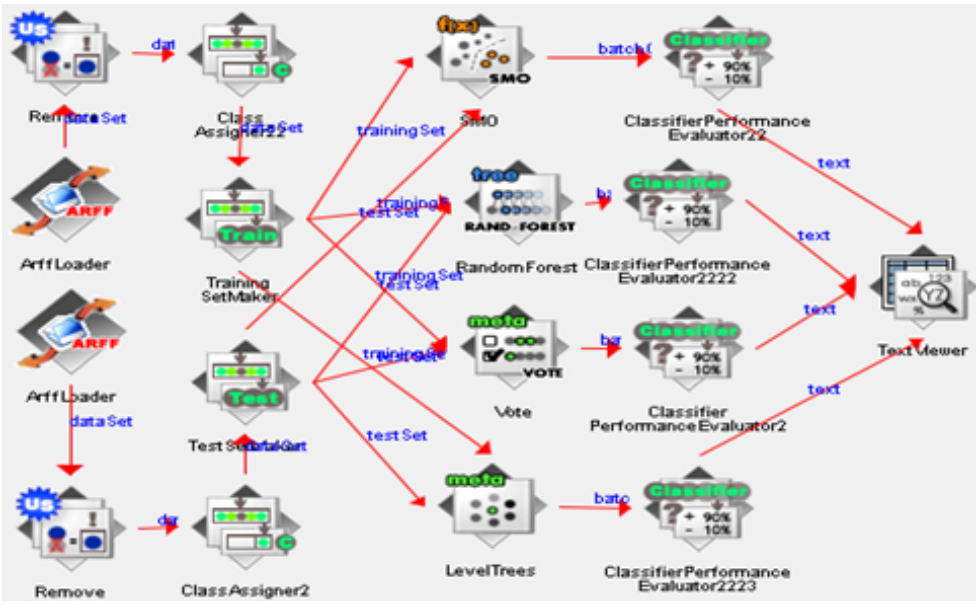

Fig. 4 – WEKA Knowledge flow voting process.

Further additional investigation on the optimization of the SMO classifier has led to an accuracy of 76.8% for the development set. The voting with average and maximum probabilities improved the best accuracy by 0.36% and 1.46%, respectively, as presented in Table 6.

Adding the Level-Trees classifier to the vote process, noted as "SMO-RF-Feating," provided the results presented in Table 7. (development and test sets).

Table 7 – Development / test sets autism classification results (Baseline accuracy : 69.8%).

	SMO		Vote(max prob.) SMO-RF		Feating		Vote : (max.prob.) SMO/RF/ Feating	
Classified Instances	devel.	test	devel.	test	devel.	test	devel.	test
Correctly	616	620	641	616	629	625	627	615
Incorrectly	203	200	178	204	190	195	192	205
Accuracy (%)	75.3	75.60*	78.26	75.12*	76.80	76.22*	76.55	75.00*
*Test results have been generated from the TUM website [7]								

5. Discussion of the results

The autism TUM-baseline [7] was developed on the basis of an SMO, with a per class up-sampling of the instances, using 6,374 attributes. The set of features was built using two framing techniques (20ms and 60ms), as presented in [7].

The TUM2013 proposed set of features, including the 60ms pitch based on the Gaussian window, the regression coefficients, and the subsequent functional coefficients, did not contribute to the autism classification. Instead, they mislead the SMO in some classes

and kept the accuracy around 69.8%, while using our proposed set of features, but by removing redundant and non-useful features, the accuracy increased by 5.80% (test set) via the SMO algorithm and by 6.42% (test set) through the feating technique.

The vote between the different classifiers improved the development results, but did not improve the test results. This is mainly due to the high similarity of the instances and the difficulties that human experts had in the manual recognition of the classes.

6. Conclusion

In this paper, we focused on a two-fold process. The first fold deals with the feature selection scheme in order to illustrate and determine the features that contribute to the autism classification, whilst the second fold concerns the vote between three different classifiers : the SMO, the RF, and the feating (Level Trees) techniques.

The final space of features decreased by 30% compared with the proposed one, with an increase of 6.42% in the classification accuracy. The vote by majority and max probability has shown good results for the SMO-Random Forest vote classifier, but decreased the overall classification by the use of the three classifiers.

The feating technique showed the best results because it is intrinsically an ensemble method, where the sublevel trees vote depends on sub-space ranked features.

The autism classification can be improved by further work on specialized autism features, and a weighted or fuzzy vote between the SMO and the feating technique.

Acknowledgements

This project was supported by NSTIP strategic technologies programs, number (12-MED2474-02) in the Kingdom of Saudi Arabia

REFERENCES

[1] National Institute of mental health, publication n°.11-5511 2011. “A Parent’s Guide to Autism Spectrum Disorder”.2011.

[2] American Psychiatric Association. “Diagnostic and Statistical Manual of Mental Disorders”, Fourth Edition - Text Revision (DSM-IV-TR).

[3] Benselama Z., Guerti M., and Bencherif M., 2007. “Arabic speech pathology therapy computer aided system,” J. of Computer Science, vol. 3, no. 9, pp. 685–692.

[4] Valentini-Botinhao C., Degenkolb-Weyers S., Maier A., Noeth E., Eysholdt U., and T. Bocklet. 2012.Automatic detection of sigmatism in children. In Proc. WOCCI, Portland, USA, pp. 1-4.

[5] Ringeval F., Demouy J., Szasz´ak G., Chetouani M., Robel L., Xavier J., Cohen D., and Plaza M., 2011., “Automatic intonation recognition for the prosodic assessment of language impaired children,” IEEE Transactions on Audio, Speech & Language Processing, vol. 19, pp. 1328–1342,

[6] Emily T. Prud’hommeaux, Brian Roark, Lois M. Black, Jan van Santen. 2011. Classification of atypical language in autism. In Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics (CMCL), 88-96.

[7] Schuller B., Steidl S., Batliner A., Vinciarelli A., Scherer K., Ringeval F., Chetouani

M., Weninger F., Eyben F., Marchi E., Mortillaro M., Salamin H., Polychroniou A., Valente F., Kim S. 2013. : "The Interspeech 2013 Computational Paralinguistics Challenge : Social Signals, Conflict, Emotion, Autism", Proc. Interspeech 2013, ISCA, Lyon, France.

[8] Tomas Pfister, Peter Robinson., 2010. "Speech emotion classification and public speaking skill assessment", Workshop on Human Behaviour Understanding, International Conference on Pattern Recognition, Istanbul, Turkey, August

[9] Breiman L., 2001. "Random forests", Machine Learning Journal, Vol., 45 :5–32,

[10] Kai Ming Ting, Jonathan R. Wells, Swee Chuan Tan, Shyh Wei Teng, 2011."Feature-subspace aggregating : ensembles for stable and unstable learners", Machine Learning Vol.82, pp :375–397,

[11] Ponti, M.P., 2011"Combining Classifiers : From the Creation of Ensembles to the Decision Fusion" Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 2011 24th SIBGRAPI, vol., no., pp. 1,10, 28-30., DOI : 10.1109/SIBGRAPI-T.2011.9.

[12] Ludmila I. Kuncheva, 2004. "Combining Pattern Classifiers : Methods and Algorithms", Wiley Editions,

[13] Florian Eyben, Martin Wöllmer, Björn Schuller : "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 25.-29.10.2010.

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009) ; The WEKA Data Mining Software : An Update ; SIGKDD Explorations, Volume 11, Issue 1.