MUTUAL INFORMATION BASED FEATURE SELECTION FOR ACOUSTIC AUTISM DIAGNOSIS

by

Şefika YÜZSEVER

B.S., Computer Engineering, İstanbul Technical University, 2007

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering
Boğaziçi University

2015

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my thesis supervisor Prof. S. Fikret Gürgen for his patience, useful comments, guidance and engagement through the learning process of this master thesis. Besides my advisor, I would like to thank Heysem Kaya who developed initiative work of my thesis, motivated me to enhance his work and gave his all technical and mental support during my study. His guidance helped me in all the time of research and writing of this thesis.

I am deeply indebted to my family, especially my son Emre Yüzsever and my husband Mert Yüzsever. Their love and patience provided me the energy to attain my study. They have been an important and indispensable source of spiritual support.

I would also want to thank my father, Rıfat Öztürk and my brothers, Bekir and Bahtinur Öztürk for keeping their support on me all the time.

Up to this stage of my life, I have always felt the support and trust of my mother in any work that I am included and in any activity that I perform. Therefore, I would specially like to thank Zeliha Öztürk, and I would like to dedicate my thesis to her.

ABSTRACT

MUTUAL INFORMATION BASED FEATURE SELECTION FOR ACOUSTIC AUTISM DIAGNOSIS

Pervasive Developmental Disorders (PDD) are known to affect children's social interactions and mental development. Prosodic and linguistic cues can be used to diagnose the disorders at early ages. Computational paralinguistics can be applied for tele-monitoring and/or educating the children with PDD. For better understanding the disorders, a small subset of highly informative features is needed. From machine learning perspective, feature selection (FS) is an important step for generalization ability of the learner and drawing inferences about the underlying problems. Since, the high dimensional data are vulnerable to comprise redundant and irrelevant features. The most popular FS methods depend on Mutual Information (MI), that resort to discretization of features. Though the effect of different discretization schemes are studied in literature, to the best of our knowledge the effect of different number of bins for equal width z-score discretization is not studied for MI based FS. Since MI computation depends on the number of discrete categories, we hypothesize that the feature ranking and therefore performance trajectory also changes. We carry out extensive experiments using eight MI based FS methods on the INTERSPEECH 2013 Autism sub-challenge corpus. The comparative results verify our hypothesis and lead to interesting remarks for future studies. Also in this thesis, adjustment for chance factor is proposed for normalizing MI measures, therefore obtaining a new MI based FS criterion. Finally, we choose the candidate ranked features by considering the effect of discretization, and achieve 70.68% Unweighted Average Recall (UAR) performance on the test set using only 2% of the feature set. This result advances state-of-the-art performance on the test set adhering to the challenge protocol.

ÖZET

AKUSTİK OTİZM TEŞHİSİ İÇİN ORTAK BİLGİYE DAYALI ÖZNİTELİK SEÇİMİ

Çocukların sosyal etkileşimi ve zeka gelişiminin yaygın gelişsel hastalıklar (YGH) tarafından etkilendiği bilinmektedir. Bu hastalıkların erken yaşta teşhis edilmesinde vezinsel ve dilbilimsel ipuçları kullanılabilir. YGH'li çocukları uzaktan izlemek ve/veya eğitmek için hesaplamasal paralinguistik uygulanabilir. Hastalıkları daha iyi anlamak için, oldukça bilgi verici özniteliklerin küçük bir altkümesine ihtiyaç vardır. Makine öğrenimi perspektifinden bakıldığında, öznitelik seçimi (ÖS) öğrenicinin genelleme kabiliyeti için ve altta yatan problemler hakkında çıkarımlar yapmak için çok önemli bir aşamadır. Çünkü, yüksek boyutlu veriler bağıntısız ve artık özniteliklerden oluşmaya eğilimlidir. Ortak bilgiye dayalı en popüler öznitelik seçim yöntemleri, özniteliklerin ayrıklaştırılmasına başvurur. Literatürde farklı ayrıklaştırma yöntemlerinin etkisi incelenmiş olmasına rağmen, bildiğimiz kadarıyla eşit genişlikte z-skor ayrıklaştırma için farklı sayıda aralığın etkisi ortak bilgiye dayalı öznitelik seçimi için çalışılmamıştır. Ortak Bilgi (OB) hesaplaması ayrık bölümlerin sayısına bağlı olduğundan, öznitelik dizimi ve dolayısıyla performans yörüngesinin değişeceğini varsaymaktayız. INTERSPEECH 2013 Otizm alt müsabaka veri kümesinde ortak bilgiye dayalı öznitelik seçim yöntemleri kullanarak kapsamlı deneyler yaptık. Karşılaştırmalı sonuçlar varsayımımızı doğrulamakta olup gelecek çalışmalar için ilgi çekici yorumlara yol açmaktadır. Ek olarak bu tezde, OB normalizasyonu için şans faktörü düzeltmesi önerilmiş ve yeni bir OB temelli ÖS kriteri elde edilmiştir. Son olarak ayrıklaştırmanın etkisini dikkate alarak aday sıralı öznitelikleri seçiyor ve özniteliklerin sadece %2'sini kullanarak test kümesinde %70.68 Ağırlıksız Ortalama Tanıma (AOT) performansı elde ediyoruz. Bu sonuç, yarışma protokülüne bağlı kalarak test kümesi üzerinde alandaki en iyi performansı iyileştiriyor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\mathbf{H}$	Hidden Output Matrix
$H(X)$	Entropy
$I(X;Y)$	Mutual Information
$I(X_i;X_j\|X_k)$	Conditional Mutual Information
$I(X_i;X_j;X_k)$	Joint Mutual Information
$R_y(a,b)$	Relevant Redundancy
S_{m-1}	Subset of Features
$\mathbf{T}$	Label Matrix
$\mathbf{W}$	Mapping Matrix
X	Random Variable
$\mathcal{X}$	Dataset
Y	Target Variable

LIST OF ACRONYMS/ABBREVIATIONS

ASD	Autistic Spectrum Disorder
AMI	Adjusted Mutual Information
AMIFS	Adjusted Mutual Information Feature Selection
ANOVA	Analysis of Variance
ASM	Acoustic Segment Model
BP	Back Propagation
CCA	Canonical Correlation Analysis
CFS	Correlation-based Feature Selection
CIFE	Conditional Informative Feature Extraction
CMIM	Conditional Mutual Information Maximization
CPSD	Child Pathological Speech Database
DNN	Deep Neural Network
EF	Equal Frequency
ELM	Extreme Learning Machine
EW	Equal Width
F0	Fundamental Frequency
FS	Feature Selection
HNR	Harmonics-to-Noise Ratio
HSD	Honest Significant Difference
JMI	Joint Mutual Information
k-NN	k-Nearest Neighbors
LLD	Low Level Descriptor
LSSVM	Least Square Support Vector Machines
maxRel	Maximum Relevance
MFCC	Mel Frequency Cepstral Coefficients
MI	Mutual Information
MIFS	Mutual Information Based Feature Selection
MIQ	Mutual Information Quotient

mRMR	Minimum Redundancy and Maximum Relevance
NMIFS	Normalized Mutual Information Feature Selection
PDD	Pervasive Developmental Disorders
PDD-NOS	PDD-not Otherwise Specified
Rasta-PLP	Rasta Style Perceptual Linear Prediction
RBF	Radial Basis Function
RF	Random Forests
RMS	Root Mean Square
RSFS	Random Subset Feature Selection
SBE	Sequential Backward Elimination
SFFS	Sequential Forward Floating Search
SFS	Sequential Forward Selection
SHS	Subharmonic Summation
SLFN	Single Layer Feed-forward Network
SLI	Specific Language Disorder
SVM	Support Vector Machines
TB	Tree Bagger
UAR	Unweighted Average Recall
WD-KNN	Weighted Discrete k-Nearest Neighbors

1. INTRODUCTION

Pervasive Developmental Disorders (PDD) are studied under different disciplines and cover a "spectrum" of developmental disorders such as Autistic Disorder, Specific Language Impairment (SLI), and PDD-Not Otherwise Specified (PDD-NOS) [1, 2]. These disorders affect the children's social interaction ability especially inhibiting proper use of language and prosody [1]. Usage of these linguistic and prosodic cues can help diagnosis and tele-monitoring of PDD. Taking care of an autistic child put emotional, financial and physical strain on parents. Early diagnosis and intervention of autism spectrum disorders can reduce the stress of the parents and improve the communication and social skills in children with autism disorders.

Detection/diagnosis of autism can be categorized under computational paralinguistics, which is the study of speakers' states (e. g., emotion, intoxication) and traits (e. g., personality, gender) apart from the spoken content. Therefore, similar signal processing and machine learning methods from related fields are applied for acoustic autism classification. In the state-of-the-art computational paralinguistics processing pipeline, the feature set is obtained by passing descriptive functionals (e. g., moments, extremes) over the acoustic Low Level Descriptor (LLD) contours (e. g., F0, shimmer, MFFC) [3]. openSMILE feature extractor [4] is commonly used to extract high dimensional (at the order of thousands) systematic features with this approach.

In machine learning literature, utilizing such a high number of features with a small amount of samples is known to reduce generalization power of the learner due to the *curse of dimensionality*. There are two main approaches for dimensionality reduction, one is feature selection which investigates a subset of original dimensions that gives the distinctive information and the other is feature extraction, which finds a new representation of original dimensions. In order to overcome curse of dimensionality and be able to explain the underlying reasons related to diagnosis, we focus on feature selection aspect in this thesis.

Feature selection (FS) methods can be broadly categorized as wrapper and filter methods [5]. Wrapper methods search a subset by means of classifier/regressor performance. Filter methods use a *heuristic merit* to drive the selection process. The wrapper methods are prone to over-fitting to data and highly depend on the choice of classifier. Filter methods are faster than wrapper methods since they do not train a classifier and the computation of subset merit is much less costly. For the case of classification, the most popular heuristic merit is based on mutual information. MI is a non-linear measure of dependence between two random variables. Many MI based methods aim to minimize feature-feature dependency (redundancy) while trying to maximize feature-target dependency (relevance) [6, 7].

The computation of MI among two continuous variables is intractable. Therefore, discretization methods are applied to make the process tractable and enhance the performance [7, 8]. Discretization can be performed in two types of methods: supervised and unsupervised, based on whether they utilize the class labels of variables in determining the breakpoints between the discrete intervals. Equal width and equal frequency discretization are popularly used unsupervised methods due to their simplicity. A commonly used approach is discretization of each feature into equal width intervals in z-normalized space [7, 8]. By its definition, MI is correlated with the number of discrete categories. Therefore the heuristic merit is affected from the number of discrete categories chosen. To the best of our knowledge, there is no study analyzing the impact of the number of discrete bins on the performance of MI based FS methods.

In this thesis, the motivation is to find the most descriptive feature subset for the acoustic autism diagnosis. The primary research focus is to apply mutual information based feature selection methods for this challenging problem. Secondary one is to compare the effect of the number of discrete intervals on a set of MI based FS methods. We utilize the corpus that is provided in INTERSPEECH 2013 Autism Sub-Challenge [9].

The remainder of this thesis is organized as follows. In Chapter 2, background on MI based FS methods, literature review and methodology are given. Chapter 3 provides experimental results, whereas Chapter 4 concludes with future directions.

2. BACKGROUND AND METHODOLOGY

In this chapter, background information about paralinguistic speech processing, information theoretic concepts, discretization, classifiers and feature reduction methods as well as related literature works are given.

2.1. Paralinguistic Speech Processing

Paralinguistics is the study of non-verbal communication that conveys emotion and nuances meaning. It deals with how the words are spoken rather than what is spoken. To understand more deeply paralinguistics, let us start with the definitions of speech and voice. In the context of paralinguistics, 'voice' is related with the acoustic properties of speaker's voice, and 'speech' is related with the spoken language with linguistics.

Voice recognition added over speech recognition gives the hints about speaker states and traits and non-verbal outbursts. For example, speaker states can be emotion, interest, health state, stress while speaker traits can be age, gender, height, personality. Non-verbal outbursts can be sighs, yawns, laughs and cries.

At the beginning, to illustrate the computational analysis of voice and speech recognition, the framework of the model is shown in Figure 2.1.

Pre-processing handles with signal properties of the speech. Speech can be consisted of multiple speakers and noise, therefore pre-processing is essential for the improvement of the speech quality. *Feature extraction* is the phase where acoustic and linguistic features are extracted. This extraction depends on the problem of the research area.

Classification/Regression deals with the categorization of the test data into either discrete or continuous targets. Classification process determines the targets such

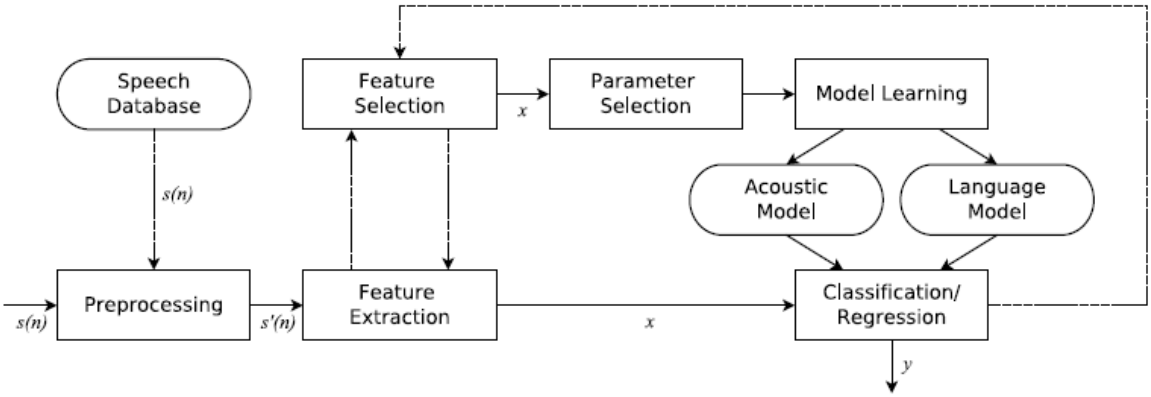

Figure 2.1. Computational speech analysis of voice and speech [10].

as emotion classes (anger, disgust, fear, happiness, sadness and surprise) or autism disorder spectrum classes. While, regression process deals with prediction of continuous value of the targets such as speaker's height in cm, age in years.

Speech database contains speech audio files for model learning and testing. Also, it may comprise the representation of the spoken content and targets such as speaker emotion, age and personality. One prefers that speech is recorded naturally and the number of speakers is as large as possible and the categorization of targets is reasonable and meaningful. *Model learning* task is performed by the classifiers/regressor, which trains the data and learns knowledge from the targets of data.

Parameter Selection refers to optimizing the parameters of the learner model. During feature selection process, speech instances for testing should not be utilized to avoid overestimation. *Acoustic Model* models learn dependencies between the acoustic features and the classes, or continuous values in the case of regression. *Language Model* is similar to acoustic models. It models learnt dependencies between linguistic features of the speech and the related targets.

Feature Selection refers to finding the most significant features for the task at hand. This step is a challenging research area of the speech analysis. For example, guessing of a speaker's age from acoustic properties is not well studied. Feature selec-

tion is the main interest of this thesis. Feature extraction phase produces an extensive number of features, however all these may not be relevant to our task. Also the feature set may contain irrelevant and redundant features. Our aim is to find an optimal feature set to improve generalization of the learner.

2.2. Literature Review

As the volume of data becomes huge in the machine learning domain, the significance of the feature selection upsurges progressively. The problem is to find the relevant features related to the task at hand. Acquisition of the efficient feature selection are the followings [11]:

- Cost of the computational time and memory is less when the feature size decreases.
- Performance of the learners enhances after omitting the redundant and irrelevant features.
- Smaller size of a feature set leads to simpler and faster learning model.
- Relevant features provides understanding about underlying data and gives intuitions.

In high dimensional data, instances appear to be sparse and diverse and this leads to over-fitting on the training data. This phenomena is known as *curse of dimensionality*. Over-fitting makes the result unreliable and inaccurate. Feature selection solves the over-fitting and accuracy problem by selecting the most discriminative and relevant features. Feature selection methods can be broadly categorized into three groups: wrapper methods, filter methods and embedded methods.

Wrapper methods aim to find a feature subset which improves the accuracy of the predictors. The number of subsets increases exponentially as the size of feature set, n increases. The exhaustive search can not be performed, even the n is about hundreds. Searching subset process may resort to heuristics for efficiency. Wrapper methods utilized from the prediction performance of the learner to evaluate the subset

relatively. Assessment of the subset is highly dependent on the given learner. The searching process is performed until the increment of the accuracy of predictor stops. Sequential Backward Elimination (SBE) and Sequential Forward Selection (SFS) are the most popular wrapper methods.

Sequential Forward Selection, proposed by Whitney [12], add features into the subset at each iteration so that the size of the subset grows progressively.

Sequential Backward Elimination, proposed by Marill and Green [13], removes features from the whole feature set progressively.

The search criteria for feature addition or removal is based on a greedy scheme. Before reaching global optimal solution, it may find a local optimal solution. Greedy search algorithms accomplish search of the solution space in a reasonable time. In SFS, selected features remain in the subset until the predefined number of feature is reached. Exclusion of a selected feature is not possible. This is known as *nesting* and it is often assumed as a disadvantage in feature selection literature since it may give rise to a local minima. Sequential Forward Floating Search (SFFS) is evolved, to overcome this issue [14]. Initially, SFFS inserts features into the subset. Then, it excludes features by backtracking until a better subset is found and continues to add new features. This algorithm combines forward selection and backward elimination and may possibly find wider range of combinations respect to SFS and SBE.

Filter methods benefit from heuristic scoring criteria to assess the importance of the features. The scoring criteria can be mutual information, Pearson correlation, Mahalanobis distance. Since filtering does not take predictors into account, it gives more generic insight about the data. Moreover, wrapper methods are prone to over-fitting than filter methods and computational time is often less in the filter methods. In a nutshell, filter methods score/rank the features according to a criteria function and select top most relevant features. The other filtering methods are feature subset selection methods [15]. Correlation-based Feature Selection (CFS) [6] and the Minimum Redundancy Maximum Relevance (mRMR) approach [7] are well known examples. CFS and

mRMR analyze correlation and mutual information, respectively. Both aim to maximize dependence between the features and the target class, while at the same time minimizing interdependence of the features in the subset. CFS determines correlation based heuristic merit between a feature set S and a target t via [6]:

$$r_{S,t} = \frac{k\overline{r_{ti}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}, \tag{2.1}$$

where k is number of features, $\overline{r_{ti}}$ denote average correlation between the features in the subset and the target variable, and the term $\overline{r_{ii}}$ denote average inter-correlation between features.

As mentioned earlier, mutual information is used for scoring criteria in feature selection. In KCCAmRMR, Sakar *et al.* [8] modifies mRMR feature selection using correlated functions of the variables (i. e., projections attained by CCA) weighted with corresponding correlations with the target class.

Currently, computational paralinguistics gains a considerable attention and motivates researchers to investigate unique and emerging systems in this area. However, lately powerful feature selection method takes the place of system design for computational paralinguistics. INTERSPEECH 2013 Challenge provides brute-forced baseline feature sets. A comprehensive literature review about feature selection and prosodic feature extraction in computational paralinguistics for autism task are the followings. The common goal of all studies is to find the most informative features for diagnosis of autism spectrum disorders. Some researchers find the relevant features using feature selection methods, some of them process the speech signal and extract their own prosodic features and combine them with baseline feature set.

In [16], the authors incorporate five subsystems for test prediction. Two of the subsystems benefit from linear kernel Support Vector Machine (SVM) using baseline features. Other two subsystems benefit from deep neural network using baseline features and the last subsystem is constructed from k-nearest neighbors (k-NN) classifier using spectral energy features. Embedding of these five systems results in 60.2% UAR.

Moreover, hierarchical classification is performed for these settings: Typical vs. Atypical; ASD vs. SLI; and PDD-NOS vs Autism. They extract 360 spectral energy related features and add MFCC and RASTA-PLP features for a total of 386 features. They implement a forward feature selection with k-NN classifier on spectral energy features. The focus of the work is to find the most discriminative prosodic features for autism spectrum disorders and specific language impairment. Effects of pitch, duration, formants, intensity, goodness of pronunciation, spectral energy and smoothness templates are examined in detail on the training and development sets.

In [17], a random subset feature selection (RSFS) is performed with k-NN classifier. The feature set is composed of the baseline features. Their study comprises three classification tasks: First one is, recognition of autism spectrum developmental disorders, second one is identifying of affective states and the last one is categorizing of level of conflict. RSFS measures relevance of each feature with respect to other features in the subset and eventually chooses all features whose relevance is over the average relevance. RSFS derives feature sets with dimensions of 430, 757 and 349 for autism, emotion and level of conflict, respectively. The selection process requires 300,000 iterations on average. Recognition of autism results in 61.9% UAR and remains below baseline UAR, 67.1%.

In [18], all sub-challenge tasks are worked out with a general machine learning meta-algorithm AdaBoost.MH [19] and AdaBoost.MH:BA [20]. Feature selection and/or extraction is not applied and the baseline feature set is used. AdaBoost integrates the base learners decisions according to their weights. The UAR scores for emotion sub-challenge are over the SVM baseline except for arousal task in the test set and also, development UAR scores outperform the SVM baseline. However, UAR score for diagnosis of autism remains below the baseline in the test set and is valued at 62.1%.

Kirchhoff *et al.* [21] developed a feature selection method based on submodular functions, which aims to find a feature subset by considering dependencies between selected features. Submodular functions optimize a general objective criteria that is

composed of similarity information between the features and the diversity information of each feature. Mutual information is utilized as a similarity measure and it evaluates pairwise dependency between two features. The study focuses on recognition of autism spectrum disorders and the classifier implemented is multi-layer perceptron. 3,000 features are outputted by submodular feature selection. On the test set, an UAR score of 64.4% is obtained, which is below the challenge baseline.

In [22], the authors combine machine learning algorithms such as SVM, deep neural networks (DNN) and weighted discrete k-nearest neighbors (WD-KNN) using baseline feature set and acoustics segment model (ASM) benefited from temporal information of the speech signal. Weights of subsystems are evaluated to decide the class output. Feature selection and/or reduction is not performed. Autism spectrum and emotion task are studied on the whole feature set. Test result of autism diagnosis in terms of UAR is 64.8%, which remains below the baseline. However, ensemble system for emotion task outperforms the baseline in the test.

In [23], the authors extract their own prosodic features and process these features to obtain iVectors and statistical descriptors. Diagnosis of autism spectrum disorder task is implemented with SVM classifier. The proposed feature set is of dimensionality 1,380. Although, classification performance of these features remains below the baseline for development set, classification performance increases when the extracted features are combined with baseline feature set in terms of UAR. On the test set, 66.06 % UAR is obtained with 6,997 features. Moreover, feature selection and/or reduction is not performed. They model pitch, energy, formants in long-term intervals, and the interval duration, shifted-delta cepstral coefficients, AM modulation index, and speaking rate to extract suprasegmental information.

In [24], the authors develop a system that extracts voice quality feature using harmonic analysis of the speech. Support vector regression is used for identifying typicality task of autism sub-challenge, and SVM is used for diagnosing autism spectrum disorders. They combine the extracted acoustic features with baseline feature set and the performance of SVM using combined features outperforms the baseline with 69.42%

UAR. Total number of features is 6,625. Moreover, feature selection and/or reduction is not performed.

The systems developed for ASD diagnosis on the INTERSPEECH 2013 Autism corpus are summarized in Table 2.1 for ease of comparison.

Table 2.1. Summary of previous studies on INTESPEECH 2013 Autism Sub-Challenge.

Work	**Features**	**Classifier**	**Test UAR(%)**
Baseline [9]	6,373	SVM	67.10
Bone *et al.* [16]	6,759	SVM,DNN and k-NN	60.20
Rasanen *et al.* [17]	430	k-NN	61.90
Goztzolya *et al.* [18]	6,373	AdaBoost.MH.BA	62.10
Kirchhoff *et al.* [21]	3,000	Multi-layer perceptrons	64.40
Lee *et al.* [22]	6,373	SVM, DNN, k-NN and ASM	64.80
Martinez *et al.* [23]	6,997	SVM	66.06
Asgari *et al.* [24]	6,625	SVM	69.42

2.3. Discretization

Pre-processing plays an important role in machine learning. It improves the performance of the classifiers and regressors. Discretization is one of the pre-processing approaches that converts continuous values into discrete ones. Effect of discretization on feature selection and classification has been studied earlier in [25, 26]. However, the approach of discretization is different from ours.

Discretization is classified into two approaches: supervised and unsupervised based on using target class labels. Equal width and equal frequency are two of the most popular unsupervised methods. *Equal Width (EW)* discretization method divides the range of a continuous variable into fixed number of intervals. Usually, the intervals are estimated by min-max normalization. The number of bins is arbitrary and depends on the dataset. However, EW discretization methods using min-max normal-

ization are susceptible to the outliers. Additionally, intervals can be estimated using z-score normalization. Here, the interval size corresponds to the standard deviation of the continuous variable. *Equal Frequency (EF)* discretization method divides the continuous variable into fixed k bins and each bin contains equal number of instance. In other words, it constructs a quasi-uniform histogram of the variable.

Supervised discretization methods utilize a broad diversity of merits based on chi-square, entropy, impurity measure and minimum description length. In this thesis, we focus on the unsupervised approach for discretization. More specifically, EW discretization after z-score normalization is used as preprocessing to expedite computation and enhance the performance of MI based feature selection methods, which are described in the next section.

2.4. Mutual Information Based Feature Selection

First of all, a brief introduction to information theoretic concepts is given. Then, the way how feature selection methods utilize entropy and mutual information concepts will be shown.

2.4.1. Entropy and Mutual Information

Entropy is a fundamental unit of information measure of random variable X, denoted by $H(X)$ measures the uncertainty on X:

$$H(X) = -\sum_{x \in X} p(x) \log p(x), \tag{2.2}$$

where the lower case x denotes a possible value of X finite sample. When X is discrete, we can easily estimate the entropy by calculating frequency counts from sample, that is $\hat{p}(x) = \frac{\#x}{N}$. N denotes the number of total samples and $\#x$ denotes the number of samples at the value x. Here, the base of logarithm determines the 'units' of the entropy. Base 2 is used in this thesis to calculate the entropy. Entropy of X reaches maximum

value when the outcome of X is unpredictable. If the possibility of all outcomes are likely equal, the uncertainty over X takes the highest value. When the distribution of X is biased to a particular outcome, the uncertainty gets lower. If the outcome of X takes only one value, the entropy is minimal. In general, $0 \leq H(X) \leq \log(|X|)$. We can define conditional entropy by using the probability equation given below. Suppose X and Y are discrete random variables, Equation 2.3 gives the entropy of X after learning Y.

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log p(x|y) \tag{2.3}$$

In other words, conditional entropy is the remaining information retrieved from X after learning outcome of Y.

In the light of these definitions, Mutual Information is the amount of information shared by random variables X and Y. Mutual dependence between X and Y is defined as,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= \sum_{x \in X} \sum_{y \in Y} p(xy) \log \frac{p(xy)}{p(x)p(y)}. \end{aligned} \tag{2.4}$$

This is the difference of the entropy of X, $H(X)$ before Y is known and the entropy of X after Y is known, $H(X|Y)$. Mutual information is symmetric, that is $I(X;Y) = I(Y;X)$. If the variables X and Y are statistically independent, MI is zero since $p(xy) = p(x)p(y)$. The relation of entropy and mutual information is shown in Figure 2.2. Likewise entropy, MI can be conditioned, as follows:

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= \sum_{z \in Z} p(z) \sum_{x \in X} \sum_{y \in Y} p(xy|z) \log \frac{p(xy|z)}{p(x|z)p(y|z)}. \end{aligned} \tag{2.5}$$

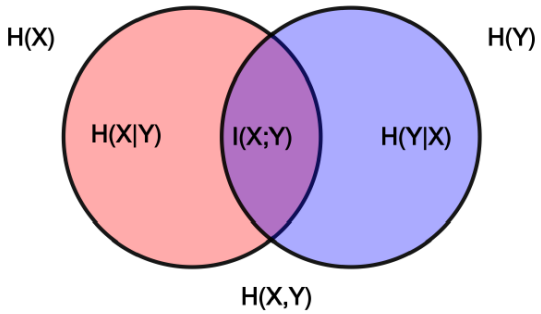

Figure 2.2. Mutual information and entropy.

This can be interpreted as the mutual information shared by X and Y after random variable Z is revealed. In the discretized case, we can estimate MI simply using the following formulation,

$$I(X;Y) = \sum_{i=1}^{K}\sum_{j=1}^{L} \frac{n_{i,j}}{N} \log \frac{n_{i,j}/N}{a_i b_j / N^2}, \tag{2.6}$$

where K and L denote the number of discrete categories for X and Y, respectively. a_i is the number of instances for i^{th} discrete variable of X (X_i), b_j is the number of instances for j^{th} discrete variable of Y (Y_j), and $n_{i,j}$ is the number of instances X_i and Y_j co-occur.

2.4.2. Mutual Information Based Feature Selection Methods

Mutual information based feature selection (MIFS) adds a feature to the subset incrementally according to their relevance and redundancy criteria in other words objective function. Relevance means dependency between feature and target class, and redundancy means dependency between two feature variables. At iteration m, the ranking method selects a feature x_m and adds to the the subset S_{m-1}. Initially subset S is empty. The first selected feature has the maximal relevance with the target class. Features may have dependency with each other hence newly added features may have redundant information about the target class.

It is important to note that the redundancy term of two discrete features $I(a;b)$ can be decomposed into relevant, $R_y(a;b)$ and irrelevant redundancy, $I(a;b|y)$ terms [8]. It can be defined as,

$$I(a;b) = R_y(a;b) + I(a;b|y), \tag{2.7}$$

where y and a, b denote the target class and features, respectively. Similarly, the following equalities hold

$$R_y(a;b) = I(a;b) - I(a;b|y) = I(y;a) - (y;a|b) = I(y;b) - (y;b|a) \tag{2.8}$$

Relevant redundancy can be represented in different ways, as seen in Figure 2.3.

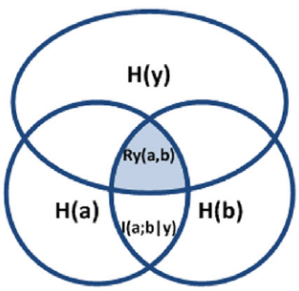

Figure 2.3. Relevant and irrelevant redundancy.

The simplest MI based feature selection method is maximum relevance (maxRel). This method aims to maximize only the mutual information between individual feature x and the target (class). Hereafter, $\mathcal{X}$ and y denote the original set and the target variable, respectively. The objective function for maxRel is defined as,

$$arg \max_{x \in \mathcal{X} - S_{m-1}} I(x;y). \tag{2.9}$$

Minimum redundancy and maximum relevance (mRMR) [7] is a first order maximal

dependency feature selection method, which optimizes the following function:

$$arg \max_{x \in \mathcal{X} - S_{m-1}} I(x;y) - \frac{1}{m-1} \sum_{i}^{m-1} I(x;x_i), \tag{2.10}$$

where first term $I(x;y)$ denotes relevance and second term expresses redundancy. Redundancy is average of the dependency between the selected feature x and features in subset S_{m-1}. Here, we see that redundancy term $I(x,x_i)$ contains both relevant and irrelevant information.

Mutual Information Quotient (MIQ) [27] is a variant of mRMR which divides relevance by redundancy. It is defined as,

$$arg \max_{x \in \mathcal{X} - S_{m-1}} I(x;y) \Big/ \frac{1}{m-1} \sum_{i}^{m-1} I(x;x_i). \tag{2.11}$$

MIQ does not follow general rule which takes the difference between relevance and redundancy into account. Therefore, in [28], the authors claim that the relevance measure is not maximized by MIQ.

Normalized mutual information feature selection (NMIFS) [29] method normalizes the redundancy between the features by minimum entropy of them. The formula of NMIFS criterion is as follows,

$$arg \max_{x \in \mathcal{X} - S_{m-1}} I(x;y) - \frac{1}{m-1} \sum_{i}^{m-1} \frac{I(x;x_i)}{min\{H(x), H(x_i)\}}. \tag{2.12}$$

Conditional Mutual Information Maximization (CMIM) [30] feature selection method subtracts only the maximum relevant redundancy between x and all x_i, as opposed to mean redundancy used in other MI based methods. Therefore, the selec-

tion criterion becomes,

$$arg \max_{x \in \mathcal{X} - S_{m-1}} I(x;y) - \max_{1<i<=m-1} R_y(x, x_i). \tag{2.13}$$

Conditional Informative Feature Extraction (CIFE) [31] method differs from CMIM in that it sums the relevant redundancy between x and all x_i:

$$arg \max_{x \in \mathcal{X} - S_{m-1}} I(x;y) - \sum_{i}^{m-1} R_y(x, x_i). \tag{2.14}$$

Joint mutual information (JMI) [32] feature selection method takes into complementary information of the candidate feature with existing features. It sums the pairwise mutual information of the candidate features and selected features. $I(x; x_i; y)$ is the mutual information between the target class and a joint feature variable x, x_i:

$$arg \max_{x \in \mathcal{X} - S_{m-1}} \sum_{i}^{m-1} I(x; x_i; y). \tag{2.15}$$

JMI can be calculated using the following formulation:

$$arg \max_{x \in \mathcal{X} - S_{m-1}} \sum_{i}^{m-1} I(x;y) + I(x_i; y|x). \tag{2.16}$$

First term denotes the information that candidate feature has related to target class and the second term denotes the information that the selected feature contains related to target class conditioned to the candidate feature.

JMI objective criterion can also be expressed in terms of relevance and redundancy with some modifications [33]. The new criteria becomes,

$$arg \max_{x \in \mathcal{X} - S_{m-1}} I(x;y) - \frac{1}{m-1} \sum_{i}^{m-1} R_y(x, x_i). \tag{2.17}$$

The two JMI criteria rank the feature set exactly in the same order, only criterion value and computational complexity differ [33].

To sum up, JMI, CMIM, and CIFE benefit from relevant redundancy, however the scaling factor of redundancy term differs in each method. mRMR and NIMFS use total redundancy which is composed of relevant and irrelevant redundancy. MIQ divides the relevance term by redundancy and this makes incomparable with other methods.

2.4.3. Proposed Method: AMI Based Feature Selection

Adjusted mutual information (AMI), which is recommend by Nguyen and Epps as a robust measure to compare clusterings in cluster analysis [34], is a variant of MI which is adjusted by expected value:

$$AMI(X,Y) = \frac{I(X;Y) - E\{I(X;Y)\}}{max\{H(X),H(Y)\} - E\{I(X;Y)\}}. \tag{2.18}$$

MI depends highly on the number of categories of the features. MI of two random variables increases as the bin size increases. Let us assume that size of the category is chosen at value nine however true number of categories is equal to five. The MI between two random variables will be higher for nine categories than the actual one. This leads to a biased result when comparing two features. Adjusted mutual information adjusts the chance factor of the MI between two random variables. Figure 2.4 illustrates the MI, normalized MI and adjusted MI of two random features from the dataset with respect to the bin size. Adjusted mutual information feature selection (AMIFS) evaluates mutual dependence as the difference of the relevance and redundancy like mRMR. The objective function is very similar to mRMR, i. e., Equation 2.10. The calculation of AMIFS differs from NMIFS [29] in adjusting both the numerator and the denominator with the expected value of MI:

$$arg \max_{x \in \mathcal{X} - S_{m-1}} I(x;y) - \frac{1}{m-1} \sum_{i}^{m-1} AMI(x;x_i). \tag{2.19}$$

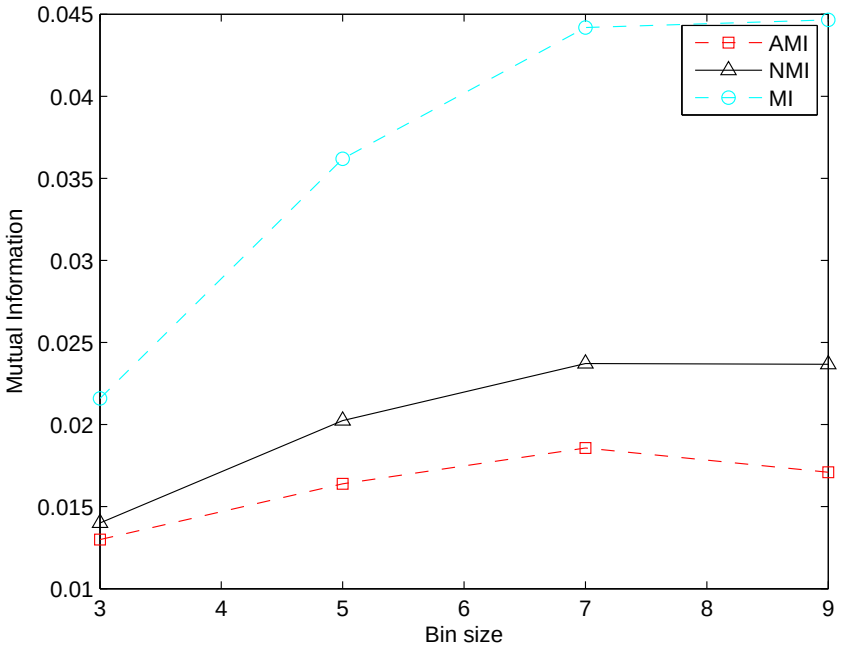

Figure 2.4. Mutual information vs bin size.

Note that while NMIFS is analogous to min-max feature normalization, AMIFS is similar to z-normalization in MI space.

2.5. Classification Methods

2.5.1. Support Vector Machines

Support Vector Machine (SVM) is a supervised learning method used for both classification and regression. SVM has been successfully applied to a wide range of pattern recognition problems such as text categorization, face detection, OCR applications and acoustic speech processing. SVM has strong mathematical foundations and it enhances performance in practical applications. The intuition of SVM is basically learning from examples.

The standard SVM is a two-class classification algorithm. SVM utilizes a maximum margin hyperplane which separates class members from non-members in the input space. SVM also constructs a nonlinear decision function in the input space by map-

ping data into a higher dimensional feature space. The algorithm categorizes a subset of informative data points called support vectors and these support vectors correspond to the hyperplane. Lastly, SVM solves a simple convex optimization problem.

Optimal Margin Hyperplane. For pattern recognition, the goal is to estimate a function $f : R^N \rightarrow \pm 1$ using training examples which are N-dimensional feature space x_i and class label y_i, indicating which class the example belongs to. For two-class classification, $y_i = +1$ or $y_i = -1$ which means examples are labeled as negative or positive. SVM constructs a hyperplane with largest margin that separates the positive from the negative examples. Thus, the hyperplane with a large margin bounds the generalization error of the classifier. The simplest model of SVM called maximal margin classifier can find a feasible solution only when the data is separable. The points X which lie on the linear separator (an optimal hyperplane) satisfy $\mathbf{wx} + b = 0$, where $\mathbf{w}$ is normal to the hyperplane and b is a threshold value. These parameters $\mathbf{w}$ and b are found by solving the following optimization problem using Lagrangian duality. The examples with non zero weights are called support vectors. On both sides of the hyperplane, instances are located $1/ \parallel \mathbf{w} \parallel$ away from the hyperplane and total margin is $2/ \parallel \mathbf{w} \parallel$ as seen in Figure 2.5.

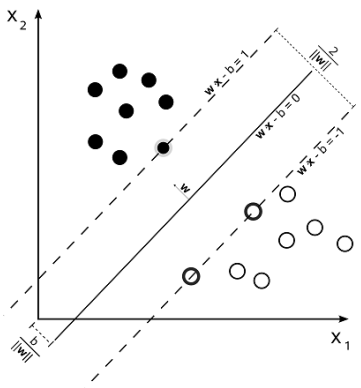

Figure 2.5. Optimal separating hyperplane.

Soft Margin Hyperplane. Another model of SVM called soft margin classifier can also find a feasible solution when data is non-separable due to the outliers and

wrong classified training examples. To avoid over-fitting, it allows some training error to some extent by introducing positive slack variables $\zeta_i = 1, ...l$ in the constraints. $\sum \zeta_i$ is an upper bound on the number of training errors.

Hyperparameters. C is a trade-off parameter that is chosen by the user. Larger C value corresponds to assignment of a higher penalty to errors. C is the regularization parameter that has an impact on generalization power of the classifier. The decision function estimates the class label of a given test example x.

Multiclass SVM. Standard SVM algorithm is a binary classifier. The straightforward method to construct multi class SVM is to reduce the single multi class problem into multiple binary classification problems. There are two common methods to construct such binary classifiers. One is one-versus-all and the other is one-versus-one. WEKA [35] uses one-versus-one method is used for multi class classification. WEKA employs sequential minimal optimization algorithm that splits the problem into a series of smallest possible sub-problems.

2.5.2. Tree Bagger

Random Forests (RF) are an ensemble of decision trees for either classification or regression. Tree Bagger (TB) is a variant of random forest that utilizes from the general technique of bootstrap aggregating, or bagging. RF generates numerous decision trees during training and these trees vote for the most popular class. Given a training set $\mathbf{X} = x_1, \ldots, x_n$ with classes $Y = y_1, \ldots, y_n$, bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples: For $b = 1, \ldots, B$:

(i) Sample, with replacement, n training examples from $\mathbf{X}$, Y; call these $\mathbf{X_b}$, Y_b.

(ii) Train a decision or regression tree f_b on X_b, Y_b.

After training, predictions for unknown samples $\mathbf{X}'$ can be obtained by taking average the predictions from all the individual regression trees on $\mathbf{X}'$:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(\mathbf{X}'), \tag{2.20}$$

or by taking the majority vote in the case of decision trees.

The bootstrapping approach enhances the generalization performance of the learner since it decreases the variance of the model without increasing bias. It aims to avoid over-fitting on the training data by using average predictions of many trees instead the predictions of a single tree which is prone to noise. Bootstrapping generates different samples so that it disables correlation of the trees by obtaining different training sets to the decision trees.

The number of the samples/trees is arbitrary. Generally, the size of trees varies from hundreds to thousands and depends on the structure of the dataset and the problem. The hyper parameter, B can be fine tuned by cross-validation. The training error settles after some numbers of trees have been fit.

RF differs from TB in the implementation of sampling process. RF uses random subset of the features during training of the model. Each decision tree trains on different feature subset of the dataset. The feature subset is sampled without replacement.

2.5.3. Extreme Learning Machines

Extreme Learning Machine (ELM) was first introduced a decade ago [36] as a fast alternative training method for Single Layer Feedforward Networks (SLFNs). The rigorous theory of the ELM paradigm is presented in 2006 by Huang *et al.* [37], where the authors compare the performance of ELM, SVM, and Back Propagation (BP) learning based SLFN in terms of training time and accuracy. The basic ELM paradigm has matured over the years to provide a unified framework for regression and classification;

and is related to generalized SLFN class including Least Square SVM (LSSVM) [38,39]. Due to fast and accurate results obtained via ELMs, the method is applied in many real life tasks ranging from gesture recognition to representational learning [40, 41]. In this section, we provide a brief introduction to the paradigm.

The argument of basic ELM introduced by Huang *et al.* is that the first layer (input layer) weights and biases of a neural network classifier do not depend on data and can be randomly generated; the second layer (output weights) can be effectively and efficiently solved via least squares [37]. It can be thought that the input layer carried out unsupervised feature mapping, then the activation function outputs (the output matrix) is subjected to a supervised learning procedure. Let $\mathbf{x} \in \mathbb{R}^d$ denote an input sample, $\mathbf{h}(\mathbf{x}) \in \mathbb{R}^p$ denote the hidden node output. Similarly, let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the dataset and $\mathbf{H} \in \mathbb{R}^{n \times p}$ denote the hidden node output matrix. The hidden node activation via randomly generated mapping matrix $\mathbf{W}$ and bias vector $\mathbf{b}$ is defined as in regular SLFN:

$$H(l,t) = h_l(\mathbf{x^t}) = g(\mathbf{x^t}, \mathbf{w_l}, b_l), l = 1, \ldots, L, t = 1, \ldots, N, \tag{2.21}$$

where l and t index the hidden nodes and the feature vectors, respectively; and nonlinear activation function $g()$ can be any infinitely differentiable bounded function [37]. A common choice for $g()$ is sigmoid function:

$$g(\mathbf{x}, \mathbf{a}, b) = \frac{1}{1 + exp\left(-(\mathbf{a} \cdot \mathbf{x} + b)\right)} \tag{2.22}$$

ELM proposes an unsupervised, even random generation of hidden node output matrix $\mathbf{H}$. The actual learning takes place in the second layer between $\mathbf{H}$ and the label matrix $\mathbf{T}$. $\mathbf{T}$ is composed of continuous annotations in case of regression therefore is a vector. In the case of K-class classification, $\mathbf{T}$ is represented in one vs. all coding

$$\mathbf{T}_{t,k} = \begin{cases} +1 & \text{if } y^t = k, \\ -1 & \text{if } y^t \neq k. \end{cases} \tag{2.23}$$

The second level weights β are learned by least squares solution to a set of linear equations $\mathbf{H}\beta = \mathbf{T}$. Proving first that random projections and nonlinear mapping with $L \leq N$ result in a full rank H, the output weights can be learned via

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \tag{2.24}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse [42] that gives not only minimum L_2 norm solution to $||\mathbf{H}\beta - \mathbf{T}||$, but also minimizes the norm of projection $||\beta||$. The use of this special generalized inverse is motivated by Barlett's theory stating that for networks approximating an arbitrarily small training error, the smaller the norm of weights is, the better the generalization capability of the network [43]. The universal approximation and classification capability of ELMs have been rigorously discussed in the literature (cf. [39]), and are beyond the scope of this thesis. However, it is important to mention that ELM is related to Least Square SVMs via the following output weight learning formulation:

$$\beta = \mathbf{H}^{T}(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^{T})^{-1}\mathbf{T}, \tag{2.25}$$

where $\mathbf{I}$ is $N \times N$ identity matrix, and C used to regularize the linear kernel, $\mathbf{H}\mathbf{H}^{T}$ is indeed the complexity parameter of LSSVM [38]. The approach is extended to use any valid kernel. A popular choice for the kernel function is Gaussian (RBF):

$$K(\mathbf{x_k}, \mathbf{x_l}) = \phi(\mathbf{x_k}) \cdot \phi(\mathbf{x_l}) = exp(-\frac{||\mathbf{x_k} - \mathbf{x_l}||}{\sigma^2}). \tag{2.26}$$

In both (basic and kernel) approaches, the prediction of $\mathbf{x}$ is given via $\hat{\mathbf{y}} = \mathbf{h}(\mathbf{x})\beta$. In case of multi-class classification, the class with maximum score in $\hat{\mathbf{y}}$ is selected. In this thesis study, we use kernel version of ELM as it was shown to outperform basic ELM in recent paralinguistic studies [44, 45].

3. EXPERIMENTS AND RESULTS

Data for acoustic autism detection are rare and usually under-sampled. Here, we use the INTERSPEECH 2013 Autism Sub-Challenge corpus, which provides both a sufficient number of instances as well as a common training, development and test platform for the competitors. In the next sections corpus and baseline feature set are introduced.

3.1. INTERSPEECH 2013 Autism Corpus

The INTERSPEECH 2013 Autism Sub-Challenge [9] uses the "Child Pathological Speech Database" (CPSD) [1]. The sub-challenge dataset comprises 2.5 k instances of speech recordings from 99 children aged 6 to 18 years. 35 subjects show autism syndrome, 29 of these are male and 6 of these are female. The control group comprises 64 subjects, 52 of these male and 12 of these female. According to DSM-IV criteria, children with autism syndrome is decomposed of three groups, Pervasive Development Disorders (PDD, 10 male, 2 female), specific language impairment such as dysphasia (DYS, 10 male, 3 female) and PDD Non-Otherwise Specified (NOS, 9 male, 1 female), that can be seen in Table 3.1. French Speech consists of the imitation of 26 sentences depicting four types of intonation and different modalities (declarative, exclamatory, interrogative, imperative). The dataset is divided into three speaker disjoint sets (training, development and test) according to order of speaker ID, age and gender. Class distribution of the dataset partitions is given in Table 3.2.

Table 3.1. Diagnosis distribution among subjects according to gender.

#	PDD	NOS	DYS	TYP
Male	10	9	10	52
Female	2	1	3	12
Total	12	10	13	64

Table 3.2. Class distribution of diagnosis according to training, devel. and test sets.

#	Train	Dev	Test
TYP	566	543	542
PDD	104	104	99
NOS	104	68	75
DYS	129	104	104
Total	903	819	820

3.2. INTERSPEECH 2013 Baseline Acoustic Feature Set

The Interspeech 2013 Challenge baseline acoustic feature set was created by modifying the acoustic features set of the previous challenge, Interspeech 2012 Speaker Trait Challenge [46]. The acoustic feature set is generated by using TUM's open-source openSMILE feature extractor. Organizers of the challenge provide extracted feature sets on a per-chunk level and a configuration file to allow for additional frame-level feature extraction. The feature set consists of 4 energy related LLD, 54 spectral LLD and 6 voicing LLD. The complete list of functionals and LLDs are given in Table 3.4 and 3.3, respectively. Totally, the previous challenge includes 6,125 features. In this challenge, they modified this feature set by improving voice quality features (jitter and shimmer), adding Viterbi smoothing for F0 and simplifying some applied functionals. Altogether, the 2013 COMPARE feature set contains 6,373 features.

Table 3.3. 65 provided low-level descriptors as given in [46].

4 energy related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
55 Spectral LLD
RASTA-style auditory spectrum, bands 1-26 (0–8 kHz)
MFCC 1–14
Spectral energy 250–650 Hz, 1 k–4 kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90
Spectral Flux, Centroid, Entropy
Skewness, Kurtosis, Variance, Slope
Psychoacoustic Sharpness, Harmonicity
6 voicing related LLD F_0 by SHS + Viterbi smoothing, Probability of voicing logarithmic HNR, Jitter (local, delta), Shimmer (local)

Table 3.4. Applied functionals. [1] : arithmetic mean of LLD / positive Δ LLD [2]: only applied to voice related LLD. [3]: not applied to voice related LLD except F_0.[4]: only applied to F_0.

Functionals applied to LLD / Δ LLD
quartiles 1–3, 3 inter-quartile ranges
1 % percentile (≈min), 99 % percentile (≈ max) position of min / max
percentile range 1 %–99%
arithmetic mean[1], root quadratic mean
contour centroid, flatness
standard deviation, skewness, kurtosis
rel. duration LLD is above / below 25 / 50 / 75 / 90 % range
rel. duration LLD is rising / falling
rel. duration LLD has positive / negative curvature[2]
gain of linear prediction (LP), LP Coefficients 1–5
mean, max, min, std. dev. of segment length[3]
Functionals applied to LLD only
mean of peak distances
standard deviation of peak distances
mean value of peaks
mean value of peaks – arithmetic mean
mean / std.dev. of rising / falling slopes
mean / std.dev. of inter maxima distances
amplitude mean of maxima / minima
amplitude range of maxima
linear regression slope, offset, quadratic error
quadratic regression a, b, offset, quadratic error
percentage of non-zero frames [4]

3.3. Experimental Results

In our experiments, we utilized Nguyen *et al.'s* [47] implementation for MI based feature selection methods[1] . We used Weka Data Mining tool [48] for Linear Kernel Support Vector Machines. We employed MATLAB [49] implementations of Tree Bagger and ELMs.

3.3.1. Feature Analysis and System Development

Firstly, up-sampling is applied using SMOTE [50] method to overcome imbalanced class distribution of the diagnosis task. The training dataset are up sampled with factor of five so that all classes become approximately equal distribution. As a preprocessing method, EW discretization is employed. As mentioned earlier, in EW discretization features are z-normalized and discretized into equal width bins. We test the effect of four different bin sizes {3,5,7,9}. Then, feature selection is carried out by mRMR, NMIFS, AMIFS, MIQ, CIFE, CMIM, maxRel, and JMI methods. Ranked features are incrementally tested with each of these FS methods via SVM. The classification simulations on ranked features are carried out using 10 to 200 features with steps of 10. Hyper-parameters of the learners, the best bin size and number of the features are tuned by cross-validation on the development set. As the performance measure, we employ Unweighted Average Recall (UAR) to counter-balance the class-imbalance. Firstly introduced in [51] as the competition measure, UAR can be defined as

$$UAR = \frac{1}{K}\sum_{k=1}^{K} TP(k)/P(k), \tag{3.1}$$

where K is the number of classes; $TP(k)$ and $P(k)$ denote the number of true positive instances and total positive instances for class k, respectively.

As a preliminary study, Tree Bagger, ELM and SVM classifiers are applied to verify the choice of classifier. Before feature selection process, data is discretized into

[1] Available from https://sites.google.com/site/vinhnguyenx/softwares

7 bins. For each feature number, hyper parameters of all classifiers are fined tuned by cross validation on the development set. The number of decision tree is set 100 for Tree Bagger. For RBF ELM, Kernel scale and complexity parameter are fined tuned in the set $10^{-5,-4,..,+3}$. SVM complexity parameter is searched in the set $10^{-4,-3,..,1}$. For ELM, min-max normalization is performed. Comparative analysis on four MI based feature selection methods are carried out. The performance of MIFS methods for each classifier on the development set is depicted in Table 3.5 .

Table 3.5. UAR(%) performance of mRMR, NMIFS, AMIFS, JMI on development set using ELM, Tree Bagger and SVM.

	RBF ELM	**Tree Bagger**	**SVM**
mRMR	62.90	54.50	61.50
NMIFS	61.70	55.30	61.50
AMIFS	62.30	55.20	61.10
JMI	60.80	54.70	58.60

All classifiers outperforms the development baseline which is 52.40% in terms of UAR. However, the performances of ELM and Tree Bagger remain below the test baseline, 67.10%. It means that ELM and Tree Bagger over-fit the data. UAR performance of RBF ELM applied with mRMR, NMIFS, AMIFS and JMI methods are 60.00%, 58.70%, 65.60% and 56.70%, respectively. Additionally, UAR performance of Tree Bagger applied with mRMR, NMIFS, AMIFS and JMI methods are 52.50%, 49.70%, 50.50% and 50.10%, respectively. However, the performance of SVM in the test set approximates the baseline result. Additionally, SVM is used as a classifier by the challenge organizers. For more comparability with test baseline and better UAR results, we decided to continue with SVM for detail feature analysis and selection. The number of selected features by mRMR, NMIFS, AMIFS, JMI is shown in Table 3.6.

To analyze the effect of discretization on MIFS comprehensively, we compare the similarity of the ranked features which are discretized by different bin sizes. For each MI method, Figure 3.1 depicts the Jaccard index, which evaluates of similarity between

Table 3.6. Number of selected features by mRMR, NMIFS, AMIFS, JMI on development set using ELM, Tree Bagger and SVM.

	RBF ELM	**Tree Bagger**	**SVM**
mRMR	120	170	140
NMIFS	200	200	170
AMIFS	110	200	150
JMI	180	190	100

ranked features with respect to the bin size. There are six combinations of different bin sizes such as 3 vs 5 bin, 3 vs 7 bin, 3 vs 9 bin, 5 vs 7 bin, 5 vs 9 bin and 7 vs 9 bin.

The Jaccard similarity coefficient measures similarity between two sets, and is defined as the size of the intersection divided by the size of the union of the sets:

$$J(A,B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \tag{3.2}$$

where A and B denote sets. Jaccard similarity is a normalized measure, i. e., becomes 1 when all the features between two subsets are the same.

From the trajectories given in Figure 3.1, we observe that the similarity increases as the bin size of pairs increases. The subsets of 7 bin size and 9 bin size are mostly the same. This indicates that mutual information computation converges with the increasing bin size. On the other hand, we see that the similarity of ranked feature sets discretized using 3 bins with those discretized using higher bin numbers are always the lowest. The dis/similarity due to bin numbers depicted in Figure 3.1 clearly shows that a use of 3 bins suggested by Peng *et al.* [7] is a naive approach.

For each feature number, SVM complexity parameter is searched in the set $10^{-4,-3,..,1}$. Then the best UAR performance for each feature number is used for comparison. We depict the performance trajectory of each method in Figure 3.2. We observe that UAR performances change with respect to bin size for all FS methods.

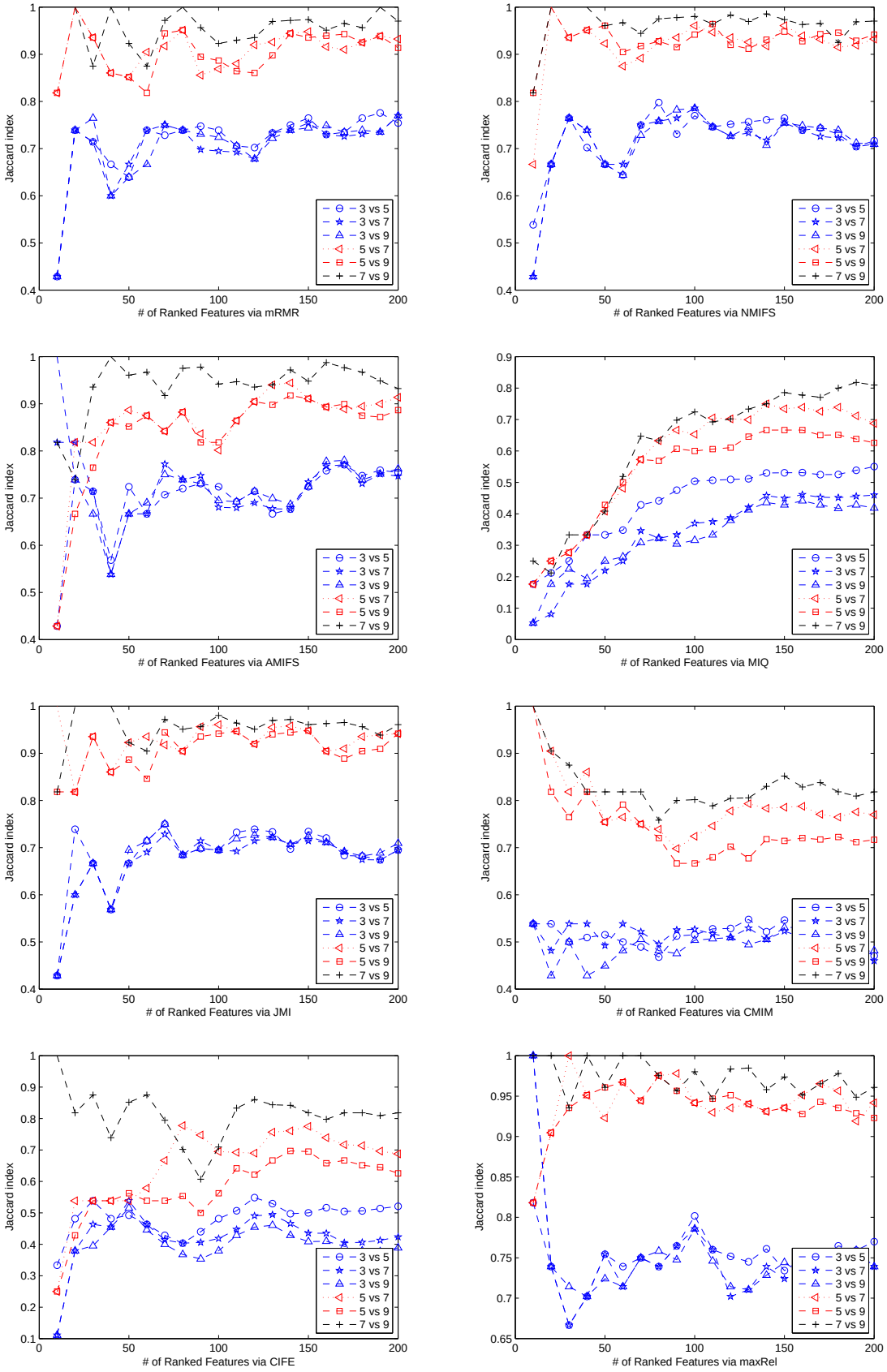

Figure 3.1. Jaccard index trajectories of eight MI based FS methods.

Though there is no single bin number that works the best for all methods, a use of 7 bins gives relatively better performance on the overall. These results verify our hypothesis that bin size affects the MI calculation. It may lead to a dramatically different ranking performance.

Figure 3.3 illustrates the performance of eight methods if the features are discretized using 7 bins, which yields the best overall performance. Though this seems a fair comparison, better results are obtainable with some other methods (e. g., CMIM and CIFE) using 3 or 5 bins discretization. The results suggest that when MI based feature selection methods are compared, using a single discretization parameter will be unfair. The performance variation due to preprocessing sometimes can be larger than the that of the method used.

The objective of this thesis study is not comparing the feature selection methods, but obtaining a necessary and sufficient subset of the features out of the candidates. Therefore, the rankings obtained from each method need to be eliminated and the number of optimal features should be determined for final prediction on the test set. To do so, we proceed with statistical analysis of ranking performance, first within each method, then between methods. The within method comparison aims to find the best or eliminating the worst number of bins, whereas between methods comparison is used to eliminate poor performing methods.

To compare the effect of bin number statistically, one-way Analysis of Variance (ANOVA) tests were applied to performance trajectories of ranked features for each method. Upon rejection of null hypothesis, post-hoc Tukey's Honest Significance Difference test (HSD) is used for pairwise comparisons. There is no significant effect of bin number on NMIFS method at the $p < 0.05$ level for the four conditions, $F(3, 76) = 2.77,\ p = 0.05$. For CMIM method, there is no significant effect of bin number at the $p < 0.05$ level, $F(3, 76) = 0.64,\ p = 0.59$. Also there is no significant effect of bin number on CIFE and maxRel methods at the $p < 0.05$ level, $F(3, 76) = 1.35,\ p = 0.26$ and $F(3, 76) = 0.95,\ p = 0.42$, respectively. In a nutshell, one-way ANOVA tests indicate no significant difference with respect to bin number

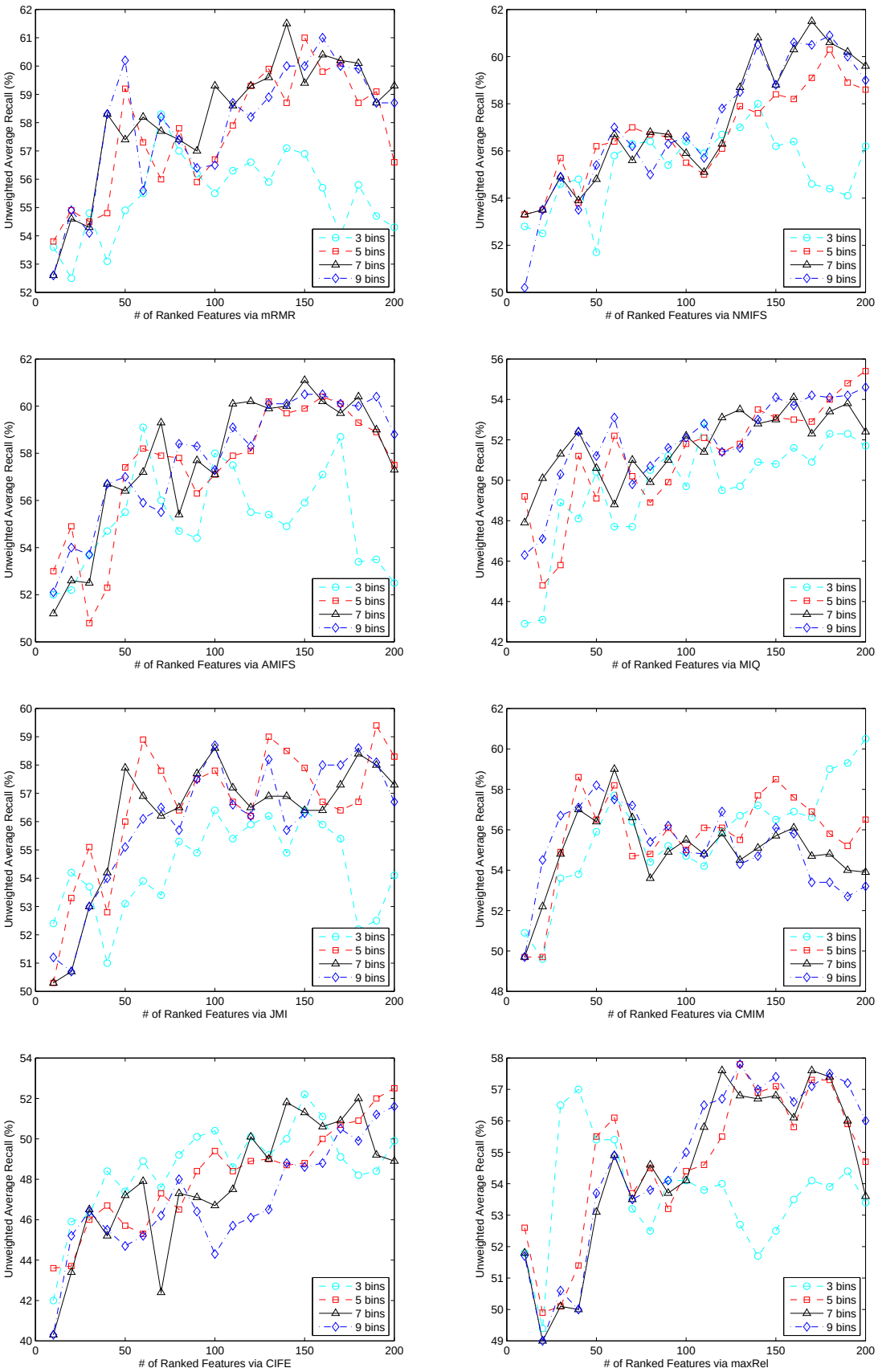

Figure 3.2. UAR performance trajectories of eight MI based FS methods with respect to varying bin sizes.

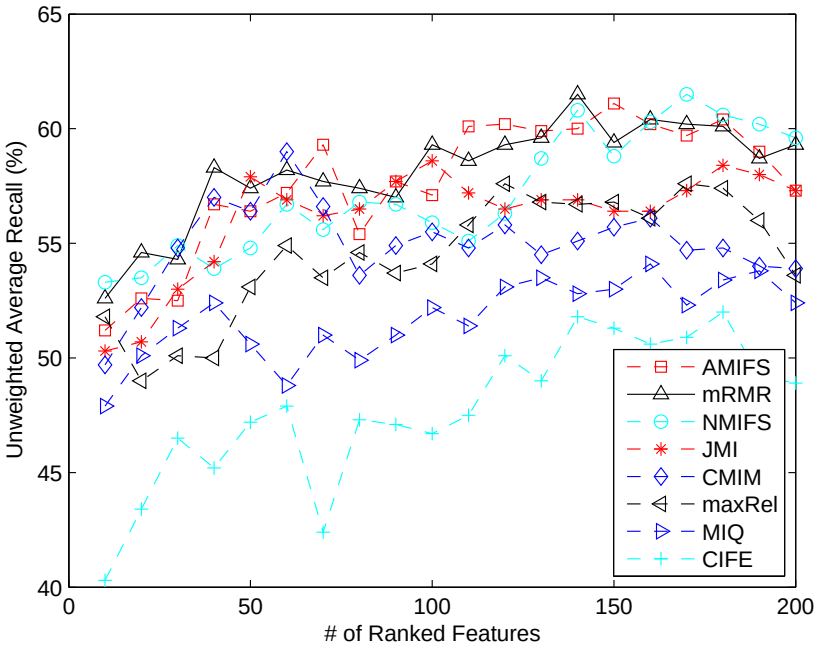

Figure 3.3. UAR performance trajectories of eight MI based FS methods using 7 bins for discretization.

in four methods, namely NMIFS, CMIM, CIFE and maxRel. For AMIFS, mRMR, JMI, MIQ methods there is a significant effect of bin number at the $p < 0.05$ level, $F(3, 76) = 4.5,\ p = 0.006,\ F(3, 76) = 7.63,\ p = 0.002,\ F(3, 76) = 4.19,\ p = 0.008$ and $F(3, 76) = 3.8,\ p = 0.014$, respectively. For these methods, post hoc tests indicate that usage of 3 bins gives the poorest results with statistically significance. While the statistical tests do not give the best bin number, the implication is to proceed the tests with highest performing bin numbers not only a single one.

Similarly, a one-way ANOVA test is applied to see whether there is a significant difference among the performances of the methods with 7 bins, which is illustrated in Figure 3.3. An analysis of variance shows that the performances of the methods differ significantly, $F(7, 136) = 34.01,\ p = 0$. Tukey's HSD test applied after rejection of null hypothesis indicate that the best performing three methods, namely NMIFS, AMIFS and mRMR have no significant difference among themselves and they are significantly better than the poorest performing four (CIFE, MIQ, maxRel and CMIM). JMI is significantly better than CIFE and MIQ, while not significantly different compared to

other better performing methods. Based on these results, the final test set predictions are carried out using NMIFS, AMIFS, mRMR and JMI with best respective bin number for each method.

Hierarchical classification is employed as a further study. Firstly, the instances are classified into two classes which are typical and atypical. Then, atypical instances are categorized into three sub groups: PDD, NOS and DYS. For each classification task feature selection is performed with numerous MI based methods using different bin sizes. Hierarchical classification is depicted in Figure 3.4.

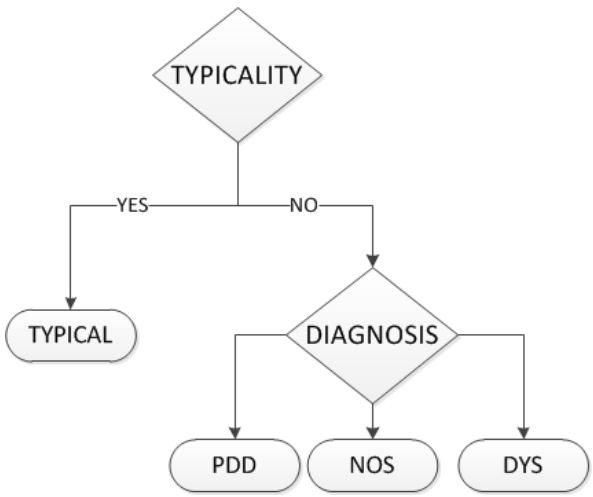

Figure 3.4. Hierarchical classification for ASD.

First step of the hierarchical classification is the typicality task. As a preprocessing, up-sampling with factor of two is performed to balance the class distribution. The best feature set is searched with 8 MI based methods with respect to bin sizes. The classification simulations on ranked features are carried out using 25 to 600 features with steps of 25. CMIM, MIQ and CIFE methods succeed better than the remaining. Their UAR performance are 93.6%, 93.0%, and 92.9%, respectively on the development set with bin size 9, 3 and 5, respectively. These results are obtained with 550 features for CMIM, 525 features for MIQ and 550 features for CIFE. The performance for typicality task of 8 FS methods on the development set is shown in Table 3.7.

Table 3.7. Performance for typicality task of 8 FS methods on the devel. set.

FS Method	**Bin Size**	**# of Features**	**UAR(%)**
CMIM	9	550	93.60
MIQ	3	525	93.00
CIFE	5	550	92.90
AMIFS	3	225	91.90
NMIFS	3	225	91.90
JMI	5	250	91.90
mRMR	3	225	91.70
maxRel	5	200	91.50

Second step of hierarchical classification is the diagnosis task. Up-sampling is not performed since the class distribution of atypical instances are approximately equal. The typical objects are omitted on the training and development set. SVM trains the model on these sets for diagnosis. CMIM and MIQ methods reach the best UAR performance 59.60% and 57.20%, respectively on the atypical development set. 70 features respect to 9 bin size for CMIM and 110 features with respect to 5 bin size for MIQ are elected. The performance for atypical diagnosis task of 8 FS methods on the development set is shown in Table 3.8. It can be concluded that for each task FS methods performs differently and number of features changes.

Overall performance of hierarchical classification is evaluated after fusing the typicality and diagnosis task. First classifier categorizes the development set into two groups: TYP and ATY. The instances which are classified as atypical are fed into second classifier. The second classifier categorizes these instances into three groups: PDD, NOS and DYS. Then, the predictions of classifiers are integrated and evaluated. The UAR performance for six combinations of typicality and diagnosis task on the development set result in 64.90%, 64.70%, 64.40%, 62.40%, 62.60%, 61.80% in descending order. The results obtained by hierarchical classification are better than the results obtained by diagnosis task on the development set. However, its performance on the

Table 3.8. Performance for atypical diagnosis task of 8 FS methods on the devel. set.

FS Method	Bin Size	# of Features	UAR (%)
CMIM	9	70	59.60
MIQ	5	110	57.20
maxRel	3	10	56.10
mRMR	3	600	55.80
AMIFS	7	40	55.80
JMI	9	550	54.70
NMIFS	5	100	54.50
CIFE	7	425	53.10

test set remains below the baseline and reaches a maximum of 62.27% UAR.

Considering the poor test set performance of the hierarchical classification, we fail to reject the null hypothesis, which assumes that hierarchical classification is not significantly better than classical classification. Thus, we focus on the regular approach for further test set predictions.

3.3.2. Challenge Test Set Results

The experiments on the development set are used to obtain sufficiently good candidate rankings and to optimize the number of bins as well as the number of features for each method. The complexity parameter of SVM to be used on the combined training and development set is not optimized on the validation set, but set to 0.001 as in the challenge baseline paper [9]. The test set performances of best candidate feature sets trained with this SVM hyper parameter are given in Table 3.9. U(S_i, S_j) means set union of selected features of corresponding systems, whereas V(S_i, S_j) means voting. We observe that AMIFS gives the best test set performance, though its performance was not the best on the development set. The achieved 70.68% UAR advances the state-of-the-art with only 2% of features used therein [24].

Table 3.9. Development and test set performances of top performing MI based methods using SVM complexity parameter $C = 0.001$.

				Devel (%)		Test (%)	
System	FS Method	#Bins	#Feats	UAR	Acc.	UAR	Acc
S1	NMIFS	9	160	53.20	69.47	69.43	82.20
S2	mRMR	7	130	53.52	69.47	67.47	81.22
S3	AMIFS	7	130	53.02	69.35	70.00	82.07
S4	JMI	5	130	52.38	68.01	69.61	82.07
S5	U(S1,S3,S4)		171	53.91	70.33	69.28	82.20
S6	V(S1,S3,S4)			53.05	69.60	70.68	82.68

Comparing these results with the previous works presented in Table 2.1, we see that our best score advances the state-of-the-art. As mentioned in Section 2.2, except the challenge winner method, all proposed methods perform below the baseline. This fact highlights the difficulty of the problem and importance of avoiding over-fitting. Note that some studies also propose feature selection methods however still use very high number of features (e. g., [21]) and fall below the challenge test set baseline.

The pairwise jaccard indices of these four systems (i. e., S1, S2, S3 and S4) are given 3.10, where we see (S2, S3) pair to have the most similar feature subset. S2 and S3 refer to mRMR and AMIFS respectively. Additionally, the percentage of the features shared by NMIFS, mRMR, AMIFS and JMI is more than 70%.

In order to analyze the interaction between acoustic signal features and autism spectrum disorders, the ranked features of best systems are categorized into four groups: Spectral, MFCC, energy related and voicing related. The distribution of selected features into major acoustic groups is shown in Figure 3.5. It is interesting to see that the proportion of voicing related features do not differ much in the selected subsets. While as expected, energy related features have higher proportion compared to their ratio in the full set. The MFCC features are originally designed for speech recognition with

Table 3.10. Jaccard index of ranked features via different MI methods on the test set.

	Jaccard Index
J(S1,S2)	0.76
J(S1,S3)	0.76
J(S1,S4)	0.75
J(S2,S3)	0.91
J(S2,S4)	0.72
J(S3,S4)	0.71

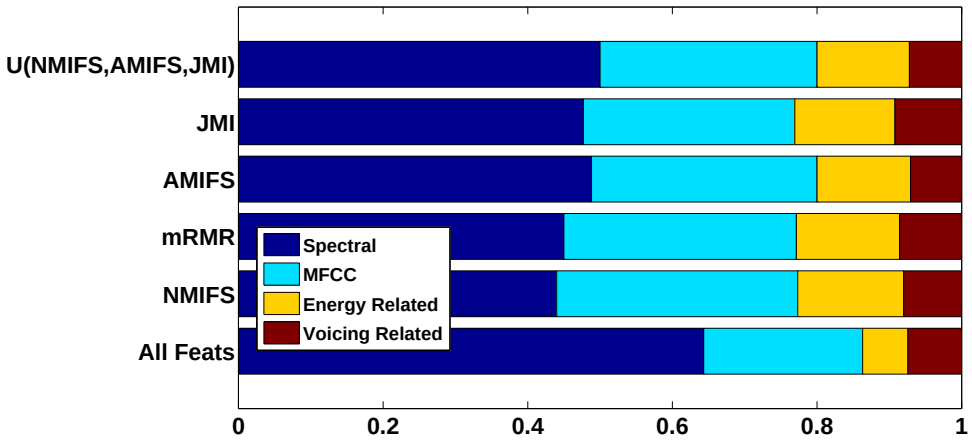

Figure 3.5. Distribution of top ranking acoustic features into major groups.

minimal effect of the speaker related factors, however they are successfully employed in a variety of tasks ranging from the gender classification to the emotion recognition. Here, we also observe that MFCC features have higher proportion in best performing subsets compared to the full set of features.

4. CONCLUSION

Autism spectrum disorders are pervasive and diverse among children. The symptoms of ASD shows variability among individuals. Therefore, clinicians spend considerable effort to diagnose ASD and their judgment may have bias. Computational paralinguistics plays an important role in tele-monitoring of these disorders and providing feedback to the clinicians. In this thesis, we analyze the acoustic signal properties of ASD by exploring the most predictive acoustic feature subsets. Thus, we better understand the interaction between ASD and speech production system.

The previous studies on this problem, including the ones that attempt to solve the problem using feature selection methods, usually failed to provide better generalization than the baseline approach on the sequestered test set. The previous results indicate the difficulty of the challenge as well as highlight the importance of generalization, i. e., avoiding over-fitting.

In this thesis, the primary focus is the application of MI based feature selection to the challenging problem of acoustic autism detection. The research sub-problem tackled thoroughly is the effect of equal width discretization after z-score normalization, which is popularly used as preprocessing in MI based feature selection methods. The contributions of the thesis are three-fold. First, we examine a set of feature selection methods, and show that the parameters used in discretization can dramatically affect the ranking. In other words, the results suggest that one size does not fit all, as long as the MI based feature ranking is concerned. Second, we propose a new feature selection criterion, AMIFS motivated from the success of NMIFS and the adjustment-for-chance property of AMI. Finally, considering the lessons learned from prior analysis on discretization, the pruned feature selection methods are applied to the INTERSPEECH 2013 Autism Sub-Challenge test set, advancing the state-of-the-art on this corpus/protocol. While the best individual results are obtained with the proposed method (UAR 70.0%), the performance is further improved with a majority voting of best three systems (UAR 70.68%).

When the features that give the highest generalization performance are analyzed, the dominance of spectral features are observed. However, proportion of the spectral features are lower compared to that in the full set. On the other hand, proportion of the energy related features and MFCC features are higher in the selected sets, compared to their prior probability obtained from the full set.

For future studies regarding MI based methods, one recommendation is the optimization of the discretization hyper-parameter (bin size) properly. The selection of this parameter depends on the data, the number of classes and the type of MI based FS method. Though the findings in this thesis are not conclusive on the optimal number of bins, the results on similarity of rankings indicate convergence of MI with respect to the bin size. Therefore, the search is not exhaustive.

Using a multi-view approach to divide-and-conquer the high-dimensional feature set, and then applying the MIFS methods on views constitute the nearest future direction. Moreover, studies on other challenging corpora are needed to further validate the virtue of the proposed feature selection method.

APPENDIX A: DETAILED RESULT TABLES

Table A.1. Best UAR (%) performance of RBF ELM of mRMR, AMIFS, NMIFS and JMI on the development set (Bin size=7).

# Feat/Meth	AMIFS	JMI	mRMR	NMIFS
10	58.4	56.7	54.8	57
20	59.1	57.9	59.2	58
30	56.2	56	57.1	56.3
40	56	55.3	56.3	55.4
50	56	57.1	58	54.4
60	56.4	57.9	56.7	55.6
70	58.4	57.1	59.1	55.8
80	57.4	56.4	58.3	56.6
90	59.3	57.6	58.1	56.2
100	60.7	57.7	59.6	56.1
110	*62.3*	58.7	60.8	57.7
120	61.8	58.1	*62.9*	57.4
130	61.6	57.9	60.7	59.6
140	61.3	58.1	61.9	59.2
150	60	57.7	60.1	57.3
160	60.3	60.1	61.7	58.2
170	61.4	60.3	61.5	57.8
180	60.4	*60.8*	62.9	59.5
190	59.8	60.3	60.6	60.7
200	59.9	59.1	62.1	*61.7*
Max	**62.3**	**60.8**	**62.9**	**61.7**

Table A.2. UAR (%) performance of Tree Bagger of mRMR, AMIFS, NMIFS and JMI on the development set(# Trees=100, Bin size=7).

# Feat/Meth	**AMIFS**	**JMI**	**mRMR**	**NMIFS**
10	47.91	46.24	45.84	46.17
20	45.34	50.35	50.00	50.67
30	50.45	48.33	48.75	50.46
40	49.65	47.62	48.74	49.51
50	50.31	48.98	49.00	49.87
60	51.86	51.66	49.20	49.36
70	52.64	51.29	52.88	50.13
80	54.53	50.30	53.12	50.03
90	54.48	52.82	54.05	51.41
100	52.29	50.81	54.32	53.17
110	55.08	53.63	53.48	52.49
120	54.11	52.73	50.01	54.74
130	51.70	54.56	52.38	53.05
140	53.66	53.69	53.80	53.82
150	53.34	52.94	53.64	52.48
160	53.68	51.49	53.57	53.14
170	51.24	53.48	*54.55*	52.92
180	55.11	53.47	53.35	53.72
190	52.31	*54.74*	54.05	53.96
200	*55.16*	54.36	52.64	*55.25*
Max	**55.16**	**54.74**	**54.55**	**55.25**

Table A.3. Best UAR (%) performance of SVM of mRMR, AMIFS, NMIFS and JMI on the development set (Bin size=7).

# Feat/Meth	**AMIFS**	**JMI**	**mRMR**	**NMIFS**
10	51.16	50.35	52.57	53.34
20	52.55	50.70	54.62	53.53
30	52.49	53.00	54.32	54.92
40	56.67	54.23	58.34	53.85
50	56.43	57.92	57.43	54.79
60	57.23	56.94	58.23	56.68
70	59.33	56.17	57.69	55.64
80	55.44	56.51	57.38	56.84
90	57.66	57.68	56.99	56.70
100	57.05	*58.63*	59.27	55.89
110	60.08	57.23	58.62	55.06
120	60.23	56.54	59.25	56.31
130	59.87	56.87	59.56	58.67
140	59.96	56.87	*61.45*	60.78
150	*61.07*	56.35	59.35	58.79
160	60.19	56.40	60.43	60.30
170	59.73	57.29	60.19	*61.48*
180	60.44	58.39	60.10	60.64
190	58.98	58.00	58.75	60.20
200	57.27	57.29	59.28	59.62
Max	**61.07**	**58.63**	**61.45**	**61.48**

Table A.4. Acoustic features ranked by NMIFS (Bin size=7).

	Acoustic Features
1	audSpec_Rfilt_sma[6]_percentile1.0
2	mfcc_sma[12]_quartile3
3	audspec_lengthL1norm_sma_quartile1
4	audSpec_Rfilt_sma[0]_flatness
5	audSpec_Rfilt_sma[3]_percentile1.0
6	mfcc_sma[9]_rqmean
7	pcm_Mag_spectralRollOff25.0_sma_percentile1.0
8	mfcc_sma_de[2]_posamean
9	audspec_lengthL1norm_sma_flatness
10	F0final_sma_flatness
11	mfcc_sma[13]_peakMeanAbs
12	audSpec_Rfilt_sma[5]_percentile1.0
13	mfcc_sma[12]_percentile99.0
14	pcm_Mag_spectralFlux_sma_flatness
15	mfcc_sma[11]_peakMeanAbs
16	audSpec_Rfilt_sma[21]_percentile1.0
17	F0final_sma_quartile1
18	audSpec_Rfilt_sma[4]_percentile1.0
19	pcm_RMSenergy_sma_de_flatness
20	pcm_Mag_harmonicity_sma_flatness
21	mfcc_sma[2]_lpgain
22	audSpec_Rfilt_sma[5]_flatness
23	audSpec_Rfilt_sma[2]_percentile1.0
24	audSpec_Rfilt_sma_de[12]_quartile3
25	F0final_sma_linregerrQ

REFERENCES

1. Ringeval, F., J. Demouy, G. Szaszak, M. Chetouani, L. Robel, J. Xavier, D. Cohen and M. Plaza, "Automatic Intonation Recognition for the Prosodic Assessment of Language-Impaired Children", *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 19, No. 5, pp. 1328–1342, 2011.

2. Demouy, J., M. Plaza, J. Xavier, F. Ringeval, M. Chetouani, D. Périsse, D. Chauvin, S. Viaux, B. Golse, D. Cohen and L. Robel, "Differential Language Markers of Pathology in Autism, Pervasive Developmental Disorder Not Otherwise Specified and Specific Language Impairment", *Research in Autism Spectrum Disorders*, Vol. 5, No. 4, pp. 1402 – 1412, 2011.

3. Schuller, B., S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load", *Proceedings of the INTERSPEECH*, pp. 427–431, 2014.

4. Eyben, F., M. Wöllmer and B. Schuller, "Opensmile: The Munich Versatile and Fast Open-source Audio Feature Extractor", *Proceeding of the International Conference on Multimedia*, pp. 1459–1462, 2010.

5. Guyon, I. and A. Elisseeff, "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, Vol. 3, pp. 1157–1182, 2003.

6. Hall, M. A., *Correlation-based Feature Selection for Machine Learning*, Ph.D. Thesis, The University of Waikato, 1999.

7. Peng, H., F. Long and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp. 1226–1238, 2005.

8. Sakar, C. O., O. Kursun and F. Gürgen, "A Feature Selection Method Based on Kernel Canonical Correlation Analysis and The Minimum Redundancy Maximum Relevance Filter Method", *Expert Systems with Applications*, Vol. 39, No. 3, pp. 3432–3437, 2012.

9. Schuller, B., S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism", *Proceedings of the INTERSPEECH*, pp. 148–152, 2013.

10. Shuller, B., "Voice and Speech Analysis in Search of States and Traits", A. A. Salah and T. Gevers (Editors), *Computer Analysis of Human Behavior*, pp. 227–253, Springer, 2011.

11. Reunanen, J., "Overfitting in Making Comparisons Between Variable Selection Methods", *Journal of Machine Learning Research*, Vol. 3, pp. 1371–1382, 2003.

12. Whitney, A. W., "A Direct Method of Nonparametric Measurement Selection", *Computers, IEEE Transactions on*, Vol. 100, No. 9, pp. 1100–1103, 1971.

13. Marill, T. and D. M. Green, "On The Effectiveness of Receptors in Recognition Systems", *Information Theory, IEEE Transactions on*, Vol. 9, No. 1, pp. 11–17, 1963.

14. Pudil, P., J. Novovičová and J. Kittler, "Floating Search Methods in Feature Selection", *Pattern Recognition Letters*, Vol. 15, No. 11, pp. 1119–1125, 1994.

15. Kohavi, R. and G. H. John, "Wrappers for Feature Subset Selection", *Artificial Intelligence*, Vol. 97, No. 1, pp. 273–324, 1997.

16. Bone, D., T. Chaspari, K. Audhkhasi, J. Gibson, A. Tsiartas, M. Van Segbroeck, M. Li, S. Lee and S. S. Narayanan, "Classifying Language-Related Developmental

Disorders from Speech Cues: the Promise and the Potential Confounds", *Proceedings of the INTERSPEECH*, pp. 182–186, 2013.

17. Räsänen, O. and J. Pohjalainen, "Random Subset Feature Selection in Automatic Recognition of Developmental Disorders, Affective States, and Level of Conflict from Speech.", *Proceedings of the INTERSPEECH*, pp. 210–214, 2013.

18. Gosztolya, G., R. Busa-Fekete and L. Tóth, "Detecting Autism, Emotions and Social Signals Using Adaboost.", *Proceedings of the INTERSPEECH*, pp. 220–224, 2013.

19. Schapire, R. E. and Y. Singer, "Improved Boosting Algorithms Using Confidence-rated Predictions", *Machine learning*, Vol. 37, No. 3, pp. 297–336, 1999.

20. Busa-Fekete, R., B. Kégl *et al.*, "Fast Boosting Using Adversarial Bandits", *Proceedings of the 27th International Conference on Machine Learning*, pp. 143–150, 2010.

21. Kirchhoff, K., Y. Liu and J. Bilmes, "Classification of Developmental Disorders from Speech Signals using Submodular Feature Selection.", *Proceedings of the INTERSPEECH*, pp. 187–190, 2013.

22. Lee, H.-Y., T.-Y. Hu, H. Jing, Y.-F. Chang, Y. Tsao, Y.-C. Kao and T.-L. Pao, "Ensemble of Machine Learning and Acoustic Segment Model Techniques for Speech Emotion and Autism Spectrum Disorders Recognition.", *Proceedings of the INTERSPEECH*, pp. 215–219, 2013.

23. González, D. M., D. Ribas, E. Lleida, A. Ortega and A. Miguel, "Suprasegmental Information Modelling for Autism Disorder Spectrum and Specific Language Impairment Classification.", *Proceedings of the INTERSPEECH*, pp. 195–199, 2013.

24. Asgari, M., A. Bayestehtashk and I. Shafran, "Robust and Accurate Features for Detecting and Diagnosing Autism Spectrum Disorders.", *Proceedings of the IN-*

TERSPEECH, pp. 191–194, 2013.

25. Ghodke, S. and T. Baldwin, "An Investigation into the Interaction Between Feature Selection and Discretization: Learning How and When to Read Numbers.", *Australian Conference on Artificial Intelligence*, Vol. 4830 of *Lecture Notes in Computer Science*, pp. 48–57, 2007.

26. Boullé, M., "Optimal Bin Number for Equal Frequency Discretizations in Supervized Learning", *Intelligent Data Analysis*, Vol. 9, No. 2, pp. 175–188, 2005.

27. Ding, C. and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data", *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, pp. 523–529, 2003.

28. Herman, G., B. Zhang, Y. Wang, G. Ye and F. Chen, "Mutual Information-based Method for Selecting Informative Feature Sets.", *Pattern Recognition*, Vol. 46, No. 12, pp. 3315–3327, 2013.

29. Estévez, P., M. Tesmer, C. Perez and J. Zurada, "Normalized Mutual Information Feature Selection", *Neural Networks, IEEE Transactions on*, Vol. 20, No. 2, pp. 189–201, 2009.

30. Fleuret, F. and I. Guyon, "Fast Binary Feature Selection with Conditional Mutual Information", *Journal of Machine Learning Research*, Vol. 5, pp. 1531–1555, 2004.

31. Lin, D. and X. Tang, "Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion.", *ECCV (1)*, Vol. 3951 of *Lecture Notes in Computer Science*, pp. 68–82, 2006.

32. Yang, H. H. and J. Moody, "Data Visualization and Feature Selection: New Algorithms for Nongaussian Data", *Neural Information Processing Systems*, pp. 687–693, 1999.

33. Brown, G., A. Pocock, M.-J. Zhao and M. Luján, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection", *Journal of Machine Learning Research*, Vol. 13, pp. 27–66, 2012.

34. Vinh, N. X. and J. Epps, "A Novel Approach for Automatic Number of Clusters Detection in Microarray Data Based on Consensus Clustering", *Bioinformatics and BioEngineering, 2009. BIBE '09. Ninth IEEE International Conference on*, pp. 84–91, 2009.

35. Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update", *SIGKDD Explor. Newsl.*, Vol. 11, No. 1, pp. 10–18, Nov. 2009.

36. Huang, G.-B., Q.-Y. Zhu and C.-K. Siew, "Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks", *Proceedings of IEEE International Joint Conference on Neural Networks*, Vol. 2, pp. 985–990, 2004.

37. Huang, G.-B., Q.-Y. Zhu and C.-K. Siew, "Extreme Learning Machine: Theory and Applications", *Neurocomputing*, Vol. 70, No. 1, pp. 489–501, 2006.

38. Suykens, J. A. and J. Vandewalle, "Least Squares Support Vector Machine Classifiers", *Neural Processing Letters*, Vol. 9, No. 3, pp. 293–300, 1999.

39. Huang, G.-B., H. Zhou, X. Ding and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification", *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, Vol. 42, No. 2, pp. 513–529, 2012.

40. Huang, G.-B., D. H. Wang and Y. Lan, "Extreme Learning Machines: A Survey", *International Journal of Machine Learning and Cybernetics*, Vol. 2, No. 2, pp. 107–122, 2011.

41. Cambria, E., G.-B. Huang, L. L. C. Kasun, H. Zhou, C.-M. Vong, J. Lin, J. Yin, Z. Cai, Q. Liu, K. Li *et al.*, "Extreme Learning Machines", *IEEE Intelligent Sys-*

tems, Vol. 28, No. 6, pp. 30–59, 2013.

42. Rao, C. R. and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*, Vol. 7, Wiley New York, 1971.

43. Bartlett, P. L., "The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than The Size of the Network", *Information Theory, IEEE Transactions on*, Vol. 44, No. 2, pp. 525–536, 1998.

44. Kaya, H. and A. A. Salah, "Combining Modality-Specific Extreme Learning Machines for Emotion Recognition in the Wild", *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 487–493, 2014.

45. Kaya, H., F. Çilli and A. A. Salah, "Ensemble CCA for Continuous Emotion Prediction", *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge*, pp. 19–26, 2014.

46. Schuller, B., S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge", *Proceedings of the INTERSPEECH*, pp. 1–4, 2012.

47. Nguyen, X. V., J. Chan, S. Romano and J. Bailey, "Effective Global Approaches for Mutual Information Based Feature Selection", *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 512–521, 2014.

48. Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations Newsletter*, Vol. 11, No. 1, pp. 10–18, 2009.

49. MATLAB, *version 7.9.0.529 (R2009b)*, The MathWorks Inc., Natick, Massachusetts, 2009.

50. Chawla, N. V., K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, 2002.

51. Schuller, B., S. Steidl and A. Batliner, "The Interspeech 2009 Emotion Challenge", *Proceedings of the INTERSPEECH*, pp. 312–315, 2009.