Journal of Computer Sciences and Applications, 2013, Vol. 1, No. 4, 61-74
Available online at http://pubs.sciepub.com/jcsa/1/4/2

DOI:10.12691/jcsa-1-4-2

Prosodic Boundary Prediction for Greek Speech Synthesis

Panagiotis Zervas*

Department of Music Technology & Acoustics, Technological Educational Institute of Crete, Rethymnon Branch, Greece
*Corresponding author: pzervas@staff.teicrete.gr

Received December 30, 2012; Revised May 18, 2013; Accepted May 19, 2013

Abstract In this article, we evaluate features and algorithms for the task of prosodic boundary prediction for Greek. For this purpose a prosodic corpus composed of generic domain text was constructed. Feature contribution was evaluated and ranked with the application of information gain ranking and correlation-based feature selection filtering methods. Resulted datasets were applied to C4.5 decision tree, one-neighbour instance based learner and Bayesian learning methods. Models performance exploitation led as to the construction of a practically optimal feature set whose prediction effectiveness was evaluated with two prosodic databases. In terms of total accuracy and F-measure, evaluation results established the decision tree effectiveness in learning rules for prosodic boundary prediction.

Keywords: *prosody, phrase breaks, ToBI, C4.5, IB1, bayesian learning*

1. Introduction

A text-to-speech (TtS) system is considered as a framework able to perform the conversion of text to synthetic speech. In this undertaking, several steps are carried out between the input information (text) and the output (synthetic speech). Macroscopically a TtS is composed of two major parts, the front-end and the back-end. Front-end accepts raw text as input and generates a symbolic representation of prosody that will be utilized for the pitch contour rendering. Finally, the back-end will process the resulted pitch contour for the generation of the synthetic waveform. Accurate construction of an appropriate pitch contour heavily depends on the utilized prosodic event description model. Extensive research led to the construction of a wide array of prosodic models examining the various prosodic events from different levels of representation; that is, acoustic level, perceptual level, and linguistic level [1]. In this article a linguistic prosodic model for the task of automatic prosodic phrasing of Greek utterances is utilized. Specifically, the adaptation of ToBI (Tone and Break Indices) [2] labelling system for Greek, the GrToBI (Greek Tone and Break Indices) [3] was utilized.

Prosodic phrasing segregates utterances into meaningful segments of information [4]. These prosodic 'chunks' occur as the speaker pauses at word junctures. Such pauses are known as prosodic phrase breaks. Since phrase breaks convey information of the spoken message, correct insertion in the appropriate word juncture is considered an important part of a TtS system. Accurate prediction of phrase breaks will affect modules of the TtS framework such as the duration module, the energy module and rendering of the pitch contour of a sentence [5]. Mistakes on this level can cause loss of naturalness and intelligibility which results alteration to the meaning of the produced sentence.

In the past, such prediction was conducted using simple phrasing algorithms [6] based on orthographic indicators, keywords or part-of-speech (POS) spotting and simple timing information. Research on the location of prosodic phrase breaks was based on the relationship of prosodic and syntactic structures. Rule-based approaches [7] applied to this particular task were most successful in applications where syntactic and semantic information were available during the generation process. Manually written rules are considered as the simplest approach of assigning prosodic phrase boundaries; even a model which simply inserts breaks after punctuation is rarely wrong, but massively underpredicts as it will allow overly long phrases when the text contains no punctuation. Moreover, complex rule driven models [8] involve much more detailed rules and require the input text to be parsed. Another weakness of this particular approach is that even if accurate syntactic and semantic information could be obtained automatically and in real time for TtS, such hand-crafted rule systems are extremely difficult to build and maintain.

Recent research on the assignment of prosodic phrase structure of text has been turned to corpus-based modelling. This approach offers the advantage of automatic construction of phrasing rules by training machine learning algorithms with large labelled corpora [9]; thus, making the adaptation to a new domain or language easier. There have been a number of models developed for the task of predicting prosodic boundaries, raging from tree-based learners [10], neural networks [11], transformational rule-based learning [12], Hidden Markov models [13], memory-based learning [14] to Bayesian learning [15].

In this paper, we evaluate features and present results of phrase break classification models constructed with the application of machine learning algorithms for Greek language. Regarding models construction, we utilized the well known C4.5 decision tree [16], one neighbour instance-based learner (IB1) [17], naive Bayes [18] and Bayesian networks [19]. Learning process was conducted with the employment of easy to extract morpho-syntactic features. Prior to learning process, we evaluated the feature effectiveness for given task by applying our data to two attribute selection approaches, the information gain ranking and the correlation based feature filtering. Attribute evaluation step led as to the construction of an optimal dataset (referred to as "practically" optimal, since it was obtained after exploiting the models performance that resulted from the feature ranking step), by excluding features with low contribution to the classification performance. Finally, two Greek prosodic databases were utilized for examining the effectiveness of the "practically" optimal feature set to the given task.

The rest of the article is organized as follows. Section 2 describes and presents details about the prosodic corpora utilized in our experiments. In section 3, the set of lexical and linguistic features extracted from our ToBI annotated data is presented and discussed. A short description of the utilized machine learning algorithms is presented in section 4. Section 5 explains the filtering methods applied to our initial dataset for the task of feature evaluation as regards Greek language. Finally, section 6 explains the structure of the process of conducting experiments and presents the results.

2. Prosodic Database Structure and Development

Extensive research in the area of speech synthesis has shown that TtS components containing quantitative models (duration module) as well as components with discrete output (such as accenting and phrasing modules) require training databases that cover effectively the output domain of an application [1]. This conclusion dictates the need of prosodic databases with adequate phonetic and prosodic coverage. Regard our data, those requirements were attained by selecting text corpora from a large amount of textual material. The initial text corpus was collected from newspaper articles and paragraphs of literature. Subsequently, the text corpus was applied as input to the letter-to-sound component producing a phoneme list as well as a diphone list. Finally, both phoneme and diphone lists were applied to the greedy algorithm [20]. Tthe acquisition of an optimal subset of the initial text corpus, containing all the Greek phonemes as well as various intra-syllabic allophones in different positions in a word structure was tge result of this endeavour.

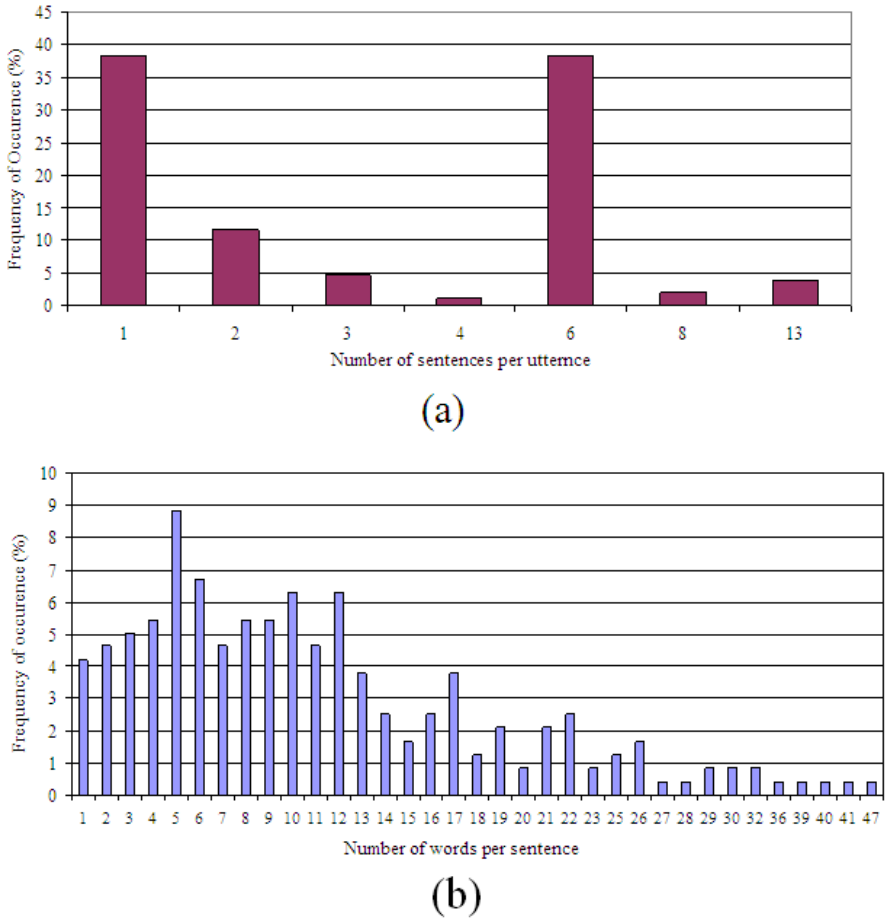


Figure 1. (a) Number of sentences per utterance (b) number of words per sentence

A major obstacle in constructing a corpus with adequate prosodic coverage is the absence of a clear definition (regarding synthesis research) of the requirements that describes it. Compared to the phonetic coverage, there is little literature talking about the requirements that should be followed, especially for Greek. Therefore, based on the

assumption that prosodic events formation is closely related to the syntactic structure of a sentence [21], we focused on the proper sentence type selection and their phrasal syntactic patterns. In dealing with cases of rare intonational and phonological phenomena, appropriate text was composed by linguists. Thus, various factors were controlled in an easier way.

The final text corpus was consisted of 5.500 words distributed in almost 494 utterances, 390 of which are declarative sentences, 44 exclamation sentences, 36 decision questions and 24 Wh-questions. Each sentence of the corpus could be a single word, a short sentence, a long sentence, or a sequence of sentences. In Figure 1a the number of sentences per utterance distribution is described while Figure 1b depicts the word number per sentence distribution. Each utterance of our text corpus could be composed of 1 to 13 sentences (with an average value of 3 sentences per utterance) while each sentence could contain from 1 to 47 words (12 words on average per utterance).

Besides sentence type, factors such as syntax, morphology, pragmatic and semantic information (Hirschberg, 1993) [22] or knowledge of "newness" and given information of the spoken message (Prevost, 1995) [23], should also be considered in order to determine the intonational pattern of an utterance. The task of extracting such information from text would require its syntactic, semantic and pragmatic analysis. Since the only information that could be examined without hand labelling were the morphological and syntactical properties of each sentence, we chose part-of-speech (POS) along with syntactic phrase boundaries as the major factors that should be considered for analysis.

2.1. Part of Speech and Syntactic Phrase Boundary Detection

POS tagging and syntactic phrase boundary detection was carried out with the application of automatic methods, followed by hand correcting the results. MG has a complex inflectional system. There are eleven different POS categories: articles (ART), nouns (N), adjectives (ADJ), pronouns (PN), verbs (V) and numerals (NUM) are declinable while adverbs (ADV), prepositions (PRE), conjunctions (CON) and particles (PRT) are indeclinable. For our purpose, we used a 2-level morphological analyzer for MG (Sgarbas et al., 1999) [24]. Figure 2 depicts the POS distribution in the final text corpus of the prosodic database.

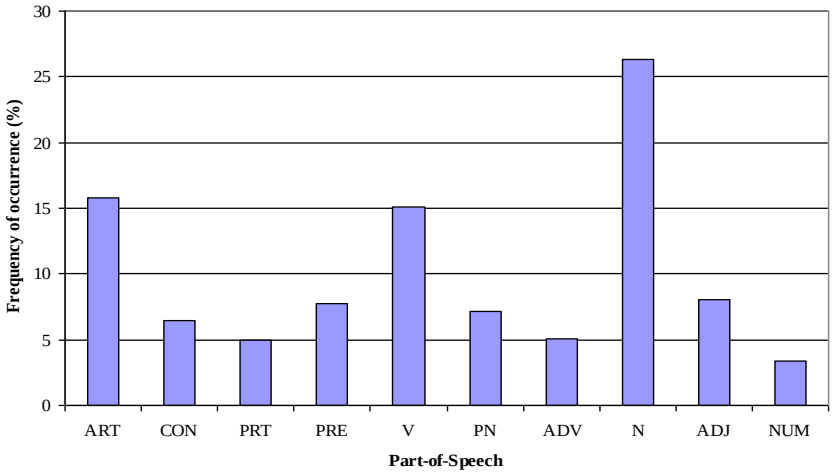


Figure 2. Part of speech distribution in the text corpus

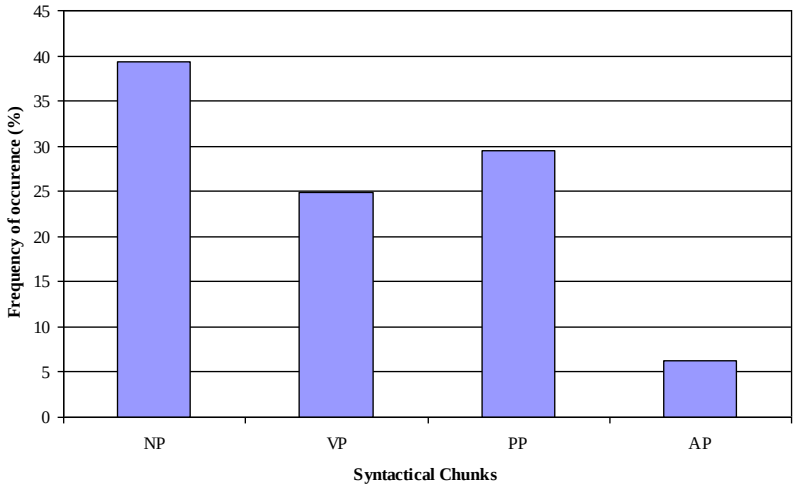


Figure 3. Syntactical phrase boundary distribution in the text corpus

The syntactic phrase boundary detector [25], or chunker, is based on very limited linguistic resources, i.e. a small function word lexicon containing some 450 keywords (articles, pronouns, auxiliary verbs, adverbs, prepositions

etc.) and a suffix lexicon of 300 of the most common word suffixes in MG. In a first stage the boundaries of non-embedded, intra-sentential noun (NP), prepositional (PP), verb (VP) and adverbial phrases (ADP) are detected via multi-pass parsing. Smaller phrases are formed in the first passes, while later passes form more complex structures. In a second stage the head-word of every phrase is identified and the phrase inherits its grammatical properties. Figure 3 depicts the distribution of the syntactical chunk categories in our database.

2.2. Speaking Style and Recording Session

Another major problem in the development of a prosodic database is the speaking style selection. Given that the main task of a TtS system is to read aloud written text, it seemed more appropriate to produce intonation of text reading. Thus, a female professional radio actress was instructed to read the selected sentences with reading style, in a normal speaking rate. A program was designed for the recording of the speech corpus. The text scripts were shown on a monitor and the recording was activated by the time the speaker started to read the sentence. The speaker was a Greek native about 30, speaking with the Athenian accent. In case of hesitations or mistakes, the speaker was asked to repeat the sentence until it was clearly pronounced. Thereby a reduction of errors in the labeling procedure could be achieved. The recording session was held in an anechoic chamber of a professional studio and took approximately 2 hours for the speaker to utter the whole text corpus. Recorded speech was sampled directly onto a DAT tape using a sampling frequency of 44.1kHz. The final data was composed of 50 minutes of clear speech, sampled onto the hard disc with a sampling frequency of 16 kHz with a resolution of 16 bit.

2.3. Prosody Annotation

As mentioned earlier description of prosody could be conducted on an acoustic, perceptual or linguistic basis. Each one of those perspectives corresponds to a different stage in the processing of prosodic information in spoken language interaction. The acoustic models of intonation include the Fujisaki's model [26], RFC [27], probabilistic models [28] and Tilt (Taylor 00) [29]. On the other hand the perceptual approach comprises the IPO model [30] and the automatic perceptual stylization model [31]. Finally, intonational models derived from linguistic analysis include the intonation theory. Since our goal was not only the reconstruction of intonational patterns, but also the exploration of effective linguistic features and the comprehension of the syntax-to-intonation relationship of Greek, we have chosen the ToBI model. Additional reasons that led as to such a decision were the following:

- ToBI is considered a standard scheme focusing on prominence and phrasing,
- designed in such a way that it is reproducible with good inter-transcribers agreement,
- and machine readable.

2.4. The GRToBI Prosody Annotation System

GRToBI encodes prosodic information for (Standard) Greek spoken corpora. In particular, it was designed for Greek as spoken in Athens. A GRToBI transcription of an utterance consists of its recording, an associated record of the pitch contour information and a file containing the GRToBI annotation tiers. The GRToBI framework is described by a five tiered annotation schema. Specifically, we have a tone tier for the intonational analysis, the prosodic words tier for phonetic transcription, a words tier for the text in Greek, a break index tier for indices of cohesion and a miscellaneous tier for other information (such as breathing, cough, etc). All the annotated information contained in the ToBI layers was aligned with time axis.

Transcribers were two linguistics graduate students and one postdoctoral researcher. The labelling of the intonational phenomena had been conducted mainly by listening to the recorded utterance in conjunction to observation of amplitude and pitch contour of the speech signal. The annotator's transcription consistency was further evaluated by cross checking statistically our data with a prosodic corpus constructed at the University of Athens for speech synthesis purposes [32].

2.4.1 The Break Index Tier

For the description of the perceived strength of each word boundary, ToBI formalization utilizes the break index tier. There are four different indices representing boundaries of different prosodic levels ranging from 0 (weaker boundary) to 3 (stronger boundary),

- Break index 0 (b0) indicates the total cohesion between orthographic words. A b0 break index denotes the presence of a single prosodic word (PrWord); co-articulation effects occur across the word boundary.
- Break index 1 (b1) marks boundaries between PrWords. Items separated by break index b1 should at most carry one pitch accent each.
- Break index 2 (b2) indicates the boundaries of an intermediate phrases (ip).
- Break index 3 (b3) denotes the boundaries of intonational phrases (IP).

Table 1 tabulates the number of occurrences of phrase break categories in our data.

Table 1. Break indices number of occurrence per word

Break Index	Number of occurrence
b0	1866
b1	2297
b2	602
b3	733

Figure 4 illustrates the correlation of break indices and punctuation as it was found in our data. We assume three levels of breaks occurred from punctuation, P0 where no punctuation existed, P2 in the case of a minor punctuation (',') and P3 for major punctuation ('.', '!', '?'). It clearly shows that b0 class is never assisted by a punctuation mark. As regards b1 class, the 91% of the occurrences are not assisted by punctuation while the rest of them are followed by minor punctuation.

Situation gets more complicated for b2 and b3 classes where both are encountered in the presence of minor or major punctuation as well as in absence. Non breaks (b0 and b1) were the most frequent categories in our prosodic database; in general, breaks (b2 and b3) are expected to be fewer than non-breaks. Since b1 category could perhaps be assigned, almost by default, between each pair of words within a sentence unless there is a punctuation mark to prevent it, high prediction results are expected. On the

other hand b3 class is encountered most of the times at the end of a sentence. This leaves the (tricky even for transcribers) question of determining a sentence-internal break to be either b2 or b3, based on the dependency relations between adjacent phrases.

Although most researchers agree that several boundary strengths must be assumed, there is no general agreement on issues such as the number and types of boundaries that need to be distinguished. In the case of prosodic phrase break prediction within TtS, it is common to flatten the prosodic; hence a word juncture is considered to be a break or a non break hierarchy [5]. In an effort to deviate from that, we considered word junctures of the entire possible phrase break label set proposed by the GRToBI transcription. Therefore, our phrase break label files contain break indices ranging from 0 to 3, where the larger number represents the end of a prosodic boundary and all the other numbers denote gradually a lower degree of decoupling.

Figure 4. Finite-state grammar for tone sequences in GRToBI

3. Features for Prosodic Phrasing Prediction

It is well established that for an accurate prediction of break indices, the extraction of textual information such as syntax and POS sequences is essential. In that way the correlation found between syntax, morphology and prosodical structure of an utterance is exploited. Since syntactic information retrieval requires both a reliable parser and a syntax-to-prosody module (which are usually implemented with the induction of rule driven methods making them complicated to write, modify, maintain and adapt to new domains and languages), we exploited syntactic phrase boundaries information along with features correlating it with the distance of adjacent syllables.

Considering the nature of the TtS synthesis challenge, only those features that can be automatically derived from text were considered. The initial feature set of our training data does not contain any attribute related to accent. We came up to this option due to the fact that prosodic phrasing is regarded as a task that precedes the prediction of accentual phenomena [5] in a TtS system. Thus, our initial feature set contains only morphological, syntactical, syllabic as well as contextual features which correlate lexical stress position, punctuation, syntactic boundaries, etc. The features utilized in our corpus are presented and described below,

- stress: whether a particular syllable is bearing a lexical stress
- syl.in: the number of syllables since last (,) or (.)
- syl.out: the number of syllables until next (,) or (.)
- ssyl.in: number of stressed syllables since last (,) or (.)
- ssyl.out: number of stressed syllables until next (,) or (.)
- last.syl.in.phrase: whether a syllable is the last in the lexical phrase or not
- last.ssyl.in.phrase: a syllable is the last stressed in the lexical phrase or not
- syl.onsetsize: number of phonemes before the vowel of a syllable
- syl.codasize: number of phonemes after the vowel of the syllable
- position.type: position of the syllable within the word
- word.numsyls: number of syllables in the word
- POS: part of speech of the word
- wrd.stress.strct: index of stress syllable in the word
- chunk: syntactic phrase boundary information
- brk.pnct: an indication of minor (',') or major punctuation ('.', '?', '!')
- chunk.in: a binary indicator showing whether a word belongs to a different syntactic chunk than its previous one
- chunk.dist: distance in words from the beginning of the next syntactic chunk or of a major punctuation break
- chunk.neighb: a binary indicator that shows whether a word belongs to the same syntactic chunk with its next one
- fc.POS: feature describing a particular word as function (FW) or content (CW)
- word.in: number of words since last (,) or (.)
- word.out: number of words until next (,) or (.)

A window of [-2, 2] to the potential boundary for each of the above features with exception to chunk.dist where a window of [-1, 1] was applied, [30]. Furthermore, to word.in, word.out, syl.in, syl.out, ssyl.in, ssyl.out, syl.codasize and syl.onsetsize no window was applied at all.

4. Prosodic Boundary Classification Framework

Several approaches for the task of automatic rule extraction from data have been developed [34] having different behaviour regarding their efficiency with certain types of class distribution than others. For our

experimental setup a set of representative learning methods for the task of phrase break prediction were employed. Thus windowed data described above, were applied to C4.5 decision tree, IB1 learner, naïve Bayes and Bayesian networks.

Decision trees have long been placed among the most practical and straightforward approaches to the task of classification [35,36]. Induction of decision trees is a method that generates approximations to discrete-valued functions with robust performance in the presence of noise. Furthermore, decision trees can be easily transformed to rules that are comprehensible by people. Decision tree classification has been applied successfully to natural language processing (NLP) tasks such as sentence boundary disambiguation [37], POS tagging [38] and syntactical parsing [39]. In the area of TtS synthesis, they have been applied for the correct placement of intonational information [40] as well as prediction of segmental durations [41].

Bayesian analysis was adduced regarding the impact certain linguistic attributes pose to the task of correctly identifying the prosodic phrase breaks by considering both the naïve Bayes and Bayesian network probabilistic assumptions. In our approach, we define a probabilistic model for resolving IP break disambiguation over a search space H*T, where H is the set of possible lexical and labelling contexts {h1,…,hk} or "variables" and T is the set of allowable phrase break labels {t1,…,tn}. There are two possible assumptions that can be considered, regarding whether the training features are considered independent of each other or taking into account a specific kind of dependency among all or a subset of them. If we assume that each lexical item is independent of all others, we adopt the naïve Bayes approach, while in the case of taking into consideration the dependency of lexical items, we apply the Bayesian networks approach

The Instance-Based (IBk) learning algorithm represents the learned knowledge simply as a collection of training cases or instances. It is a form of supervised learning from instances; it keeps a full memory of training occurrences and classifying new cases using the most similar training instances. A new case is then classified by finding the instance with the highest similarity and using its class as prediction. IBk algorithm is characterized by a very low training effort. This leads to high storage demands caused by the need of keeping all training cases in memory. Furthermore, one has to compare new cases with all existing instances, which results in a high computation cost for classification. After an extensive number of experiments we concluded to the utilization of IBk for k=1 (one neighbour). All algorithms were acquired from the WEKA machine learning library [34].

5. Feature Evaluation

The majority of machine learning algorithms are designed to decipher the most appropriate features and to utilize them for carrying out their decision. Decision tree methods, for example, choose the most promising attribute to split on at each point and, theoretically, never select irrelevant or unsupportive attributes. Thus, the higher the number of features the more discriminating power of the classifier; which is not correct since adding irrelevant or distracting attributes to a dataset often perplexes machine learning systems. Furthermore, decision tree classification performance is affected dearly with the addition of a random binary attribute, causing it to deteriorate. Thus, during decision tree's learning process an inappropriate attribute is always chosen to branch on, causing random errors during evaluation process. As you decent further down the tree structure, less data is available to assist the selection decision. Meaning that, at a certain point of the training procedure you inevitably reach depths at which only a small amount of data is available for attribute selection. When training is carried out with large datasets it would not necessarily help an attribute selection procedure; since you would possibly grow a larger tree. However in the case of small training datasets, as ours, attribute selection step is considered essential.

Divide-and-conquer tree learners and separate-and-conquer rule learners both suffer from this effect for the reason that they inexorably reduce the amount of data on which they base judgments. As regards instance-based learners, they are very susceptible to irrelevant attributes as they always work in local neighbourhoods, taking a few training instances into account for each decision. It has been shown that the number of training instances needed to produce a pretender-mined level of performance for instance-based learning increases exponentially with the number of irrelevant attributes present [42]. Finally, a classifier like naive Bayes which assumes by design, that all attributes are independent of one another, is also affected by irrelevant attributes since its operation is damaged by their presence. All the above establish the necessity of an attribute filtering step to our classification framework since it, reduces the dimensionality of the data by deleting unsuitable attributes and improves the performance of learning algorithms and presents knowledge regarding the contribution of each feature for the task of phrase break classification.

Algorithms that perform feature selection as a pre-processing step prior to learning can generally be placed into one of two broad categories. One approach referred to as the wrapper [43] employs a statistical re-sampling technique (such as cross validation) using the actual target learning algorithm to estimate the accuracy of feature subsets. This approach has proved useful but it is very slow to execute because the learning algorithm is called repeatedly. Another approach called the filter [44] operates independently of any learning algorithm - undesirable features are filtered out of the data before induction commences.

For our experiments we selected to exploit two well established approaches for feature evaluation, the Information Gain (IG) approach and the Correlation-Based feature selection (CFS) [45]. Both attribute selection methods belong to the filter category. IG was selected since, with the application of ranker method, produces the ranking of all features in the dataset based to their contribution to the classification of the desired category. On the other hand CFS selection was selected since it evaluates the worth of feature subsets of a given dataset. It has been shown [45] that CFS performance compares favourably with the wrapper but requires much less computation. Both feature selection approaches were not performed on the full dataset; instead 10 fold cross validation [46] was utilized.

5.1. Information Gain Feature Ranking

Table 2 tabulates the IG ranking of our initial feature set of the prosodical database (where pp.means previous previous, p.means previous, n.means next and nn.means next next for a [-2, 2] window). The analysis of Table 2 data, verified that phrase break class is highly correlated with almost every feature containing knowledge of lexical phrasing. Specifically, lexical punctuation (brk.pnct) showed the highest IG, followed by word.out, in contrast to word.in which had a low position to the ranking table. Attributes representing knowledge of POS, function/content word distinction, or syntactical phrasing identity of the word (chunk) also benefited the classification task. It is important to emphasize the fact that the introduced features combining syntactical phrasing identity (chunk) with its position to the sentence structure (chunk.neighb, chunk.in, chunk.dist) showed higher IG than the chunk attribute itself. Features conveying morphological information such as word stress structure (word.stress.strct), number of syllables of a word (word.numsyls) was highly correlated to the prosodic boundary class. On the other hand many of the phonological (syllabic) features were not used at all. The resulted ranking of features for Greek validates the observation of previous works in several languages claiming that prosodic boundaries prediction is strongly connected to the morpho-syntactic structure of the utterance [5].

Table 2. Feature information gain ranking

Features	pp	p	C	n	nn
brk.pnct	23	11	1	7	13
word.out	2				
POS	44	34	3	21	14
word.stress.strct	55	56	4	25	19
word.numsyls	49	47	5	24	17
position.type	50	22	6	12	15
syl.out	8				
fc.POS	59	57	9	28	20
ssyl.out	10				
chunk.neighb	70	54	16	31	33
last.syl.in.phrase	68	62	27	29	42
last.ssyl.in.phrase	61	36	30	48	65
chunk.dist	45	35	32		
syl.in	38				
word.in	40				
ssyl.in	41				
chunk	53	46	43	26	18
chunk.in	66	60	51	37	39
sStress	69	67	52	71	63
syl.codasize	58				
syl.onsetsize	64				

5.2. Correlation Based Feature Subset

Filtering of a given feature set with CFS is carried out by taking into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them. In specific, it assumes that an optimal feature subset should contain features highly correlated with the class, yet uncorrelated with each other. Initially, feature-class and feature-feature correlations are calculated with the employment of symmetrical uncertainty followed by the searching of feature subset space. The subset with the highest relevance to the class is used to reduce the dimensionality of both the original training data and the testing data. Both reduced datasets may then be passed to a machine learning algorithm for training and testing.

Application of CFS filter to our dataset resulted a feature subset constituted of, p.brk.pnct, brk.pnct, n.brk.pnct, POS, word.stress.strct, position_type, fc.POS, chunk.neighb. The fact that certain features achieved a high ranking position in IG filtering and were not selected by CFS, was due to their high correlation with other features that were already selected by the selection procedure since they were more connected to the class.

6. Experimental Framework

The evaluation schema followed in this work is composed of three parts. Initially, based on IG feature ranking results, datasets were built in the following manner; the first dataset contained only the first feature of IG ranking (that is c.brk.pnct feature, Table 2), the second dataset was composed of the previous dataset plus the next feature with highest IG (that is c.word.out). Following that pattern and by adding the next feature in the IG ranking to the former dataset, we would be able to have an insight of feature efficiency to the given task by taking into account its correlation with the previous features.

The second part in our experimental framework was the construction of phrase break models by training the selected machine learning classifiers with the CFS subset. Finally, in the third part, construction and evaluation of the "practically" optimal dataset was performed. This dataset was resulted from the initial feature repository by excluding attributes having a negative contribution to prediction's total accuracy. The contribution was based on the experiments carried out with the IG ranked datasets. The "practically" optimal dataset efficiency was evaluated

with experiments on our prosodic database as well as on a limited domain database previously utilized for prosodic modelling of Greek speech.

Performance of the resulted prosodic boundary prediction models was measured with the employment of total accuracy and F-measure per class. F-measure metric is defined as the harmonic mean of precision and recall. All boundary prediction models were evaluated with the utilization of 10 fold cross validation methodology.

6.1. IG Ranked Feature Datasets Evaluation

Figure 5 illustrates the total accuracy of C4.5, IB1, Naïve Bayes and Bayesian network prosodic phrase break models trained with datasets resulted from the IG filtering step. It is clearly shown that C4.5 results models with higher total accuracy compared to those acquired with the other classification algorithms. Specifically, C4.5 models achieved a mean total accuracy of 85.56% while IB1 had 76%, naive Bayes 74.67% and Bayesian Networks 77.12%.

C4.5 models total accuracy seems more stable, compared to the other learning schemas, in the addition of ranked features. This can be explained by the detail that, during C4.5 tree growing procedure less relevant features to the classification category, are used to nodes residing lower to the tree structure. Thus, superior discrimination capability compared to the other algorithms for the IG ranked datasets was achieved.

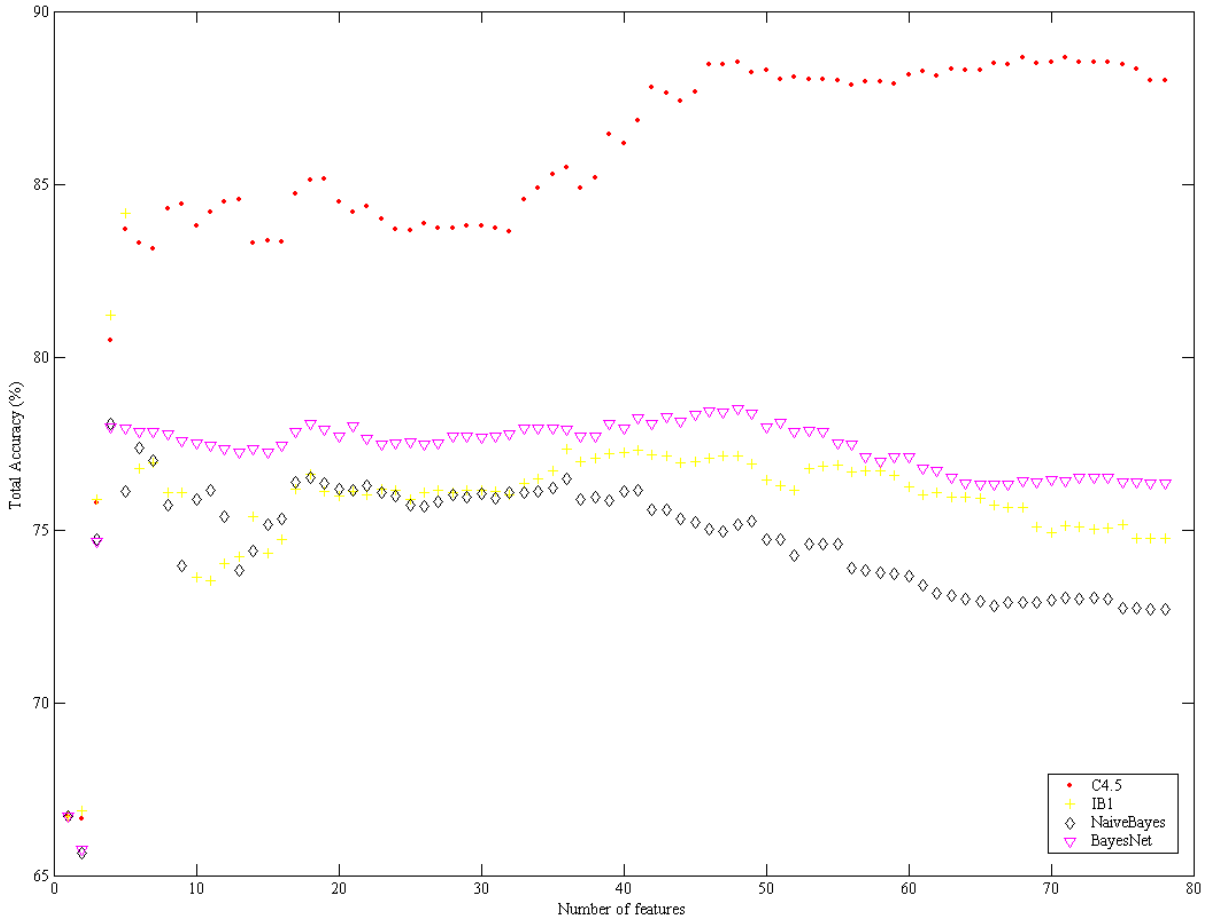


Figure 5. Total accuracy of learning models trained with IG ranked datasets.

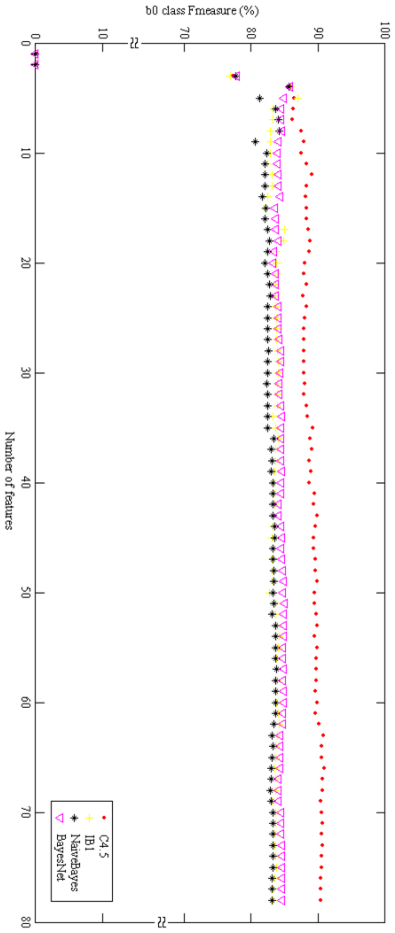
In order to posses a better comprehension of each model's performance concerning phrase break prediction, the F-measure scores achieved for each class are presented in Figures 6 a, b, c and d. Assumptions made in section 2.4.1 regarding the non-break and break classes, are clearly displayed in Figure 6. In specific, Figures 6.a and 6.b which illustrate the F-measure scores of the non-break classes, shows that both were robustly predicted with a mean F-measure score, for all training datasets and learning methods, of 82% and 83% respectively. C4.5 had the highest F-measure score for both classes, with a max value of 90.8% for b0 and 91,4% for b1 among all IG filtered datasets. Figures 6.c and 6.d present the F-measure scores of break classes, b2 and b3. For these categories, C4.5 showed a maximum F-measure of 72%, IB1 59%, Naïve Bayes 45,4% and Bayesian networks 50%. A closer inspection of Figure 6.c reveals that prediction of b2 category was enhanced greatly with the addition of word.out, word.stress.strct, chunk.neighb, word.numsyls and syl.in for all learning schemas.

6.2. CFS Subset Evaluation Results

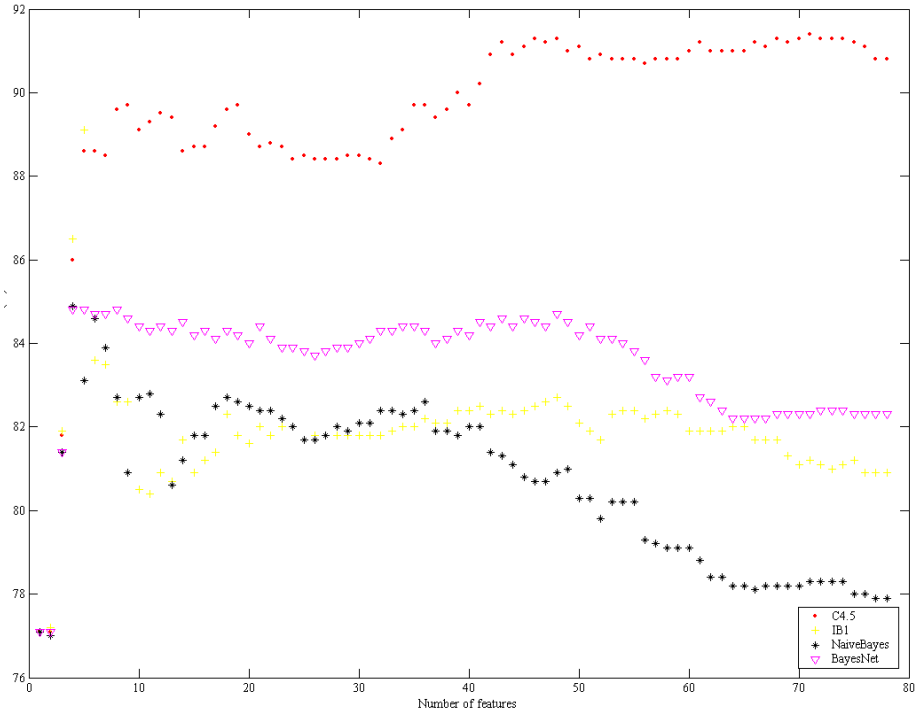
The second part of our feature and algorithm evaluation describes the experiments carried out with the CFS subset. As explained in section 5.2, the CFS procedure produces a minimal subset of attributes that are highly correlated to the predicted class. The total accuracy scores of the models resulted from CFS subset training were 86.38% for C4.5, 85.95% for IB1, 75.65% for naive Bayes and 77.21% for the Bayesian networks.

For this subset of features all algorithms performed equally well as regards the prediction of b0, b1 and b3

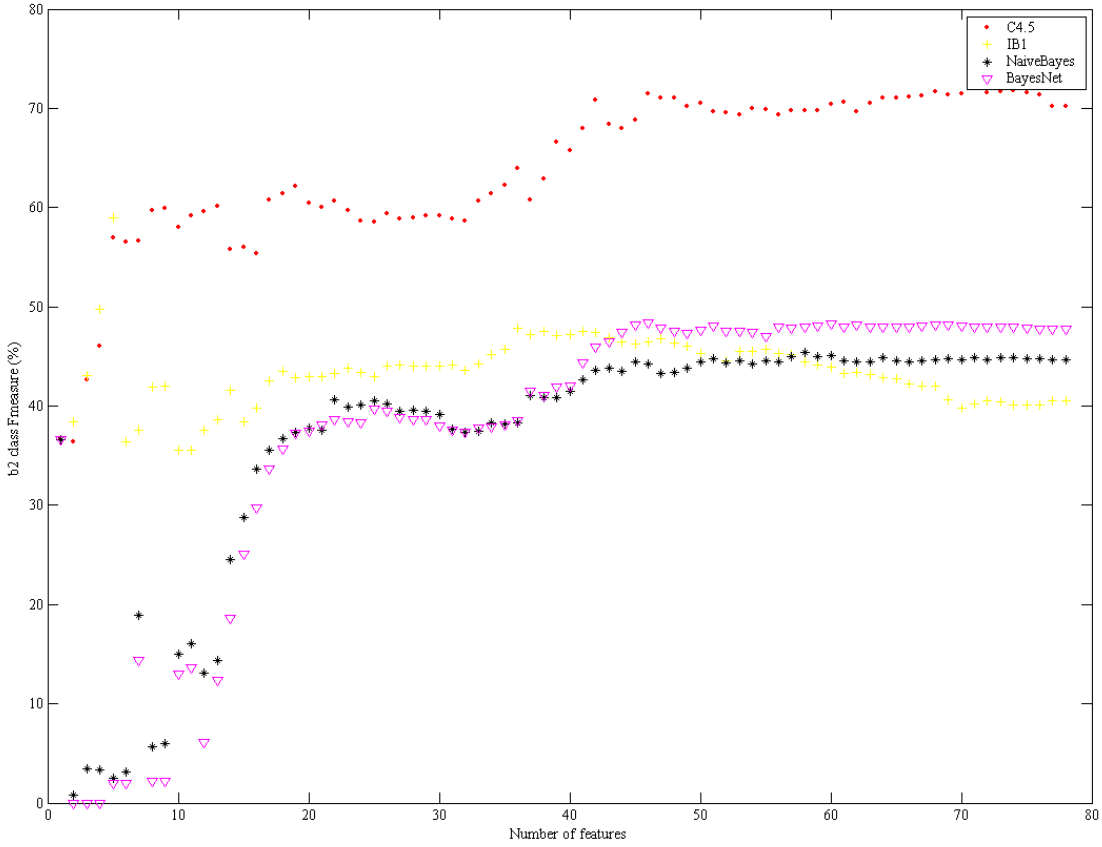
classes. Furthermore, C4.5 and IB1 outperformed naive Bayes and Bayesian networks for the prediction of the b2 category. Figure 7 depicts that C4.5 and IB1 performed better for all phrase breaks categories compared to naive Bayes and Bayesian networks. In specific, for the prediction of b2 category, C4.5 and IB1 outperformed naive Bayes and Bayesian networks models. F-measure score achieved for the prediction of b2 category was, 67.8% and 67.6 % for C4.5 and IB1 while naive Bayes and Bayesian networks achieved 31.2% and 30% respectively.



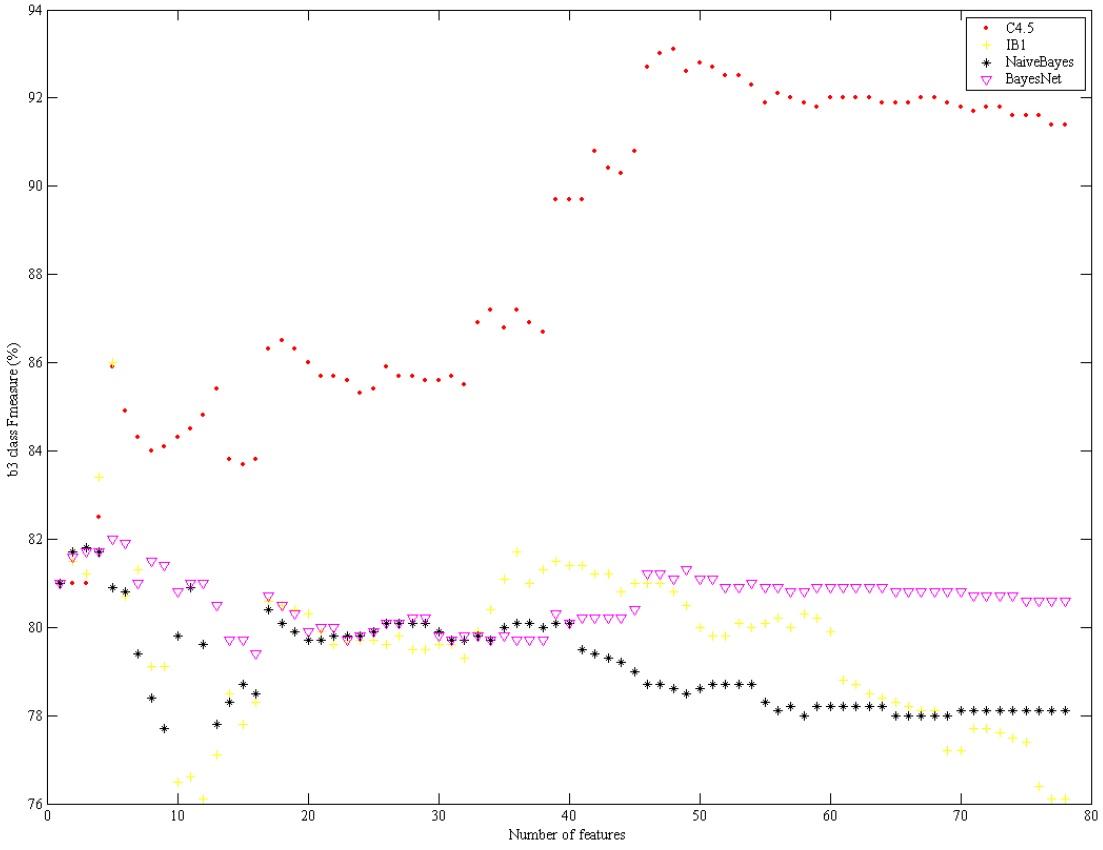
(a)



(b)



(c)



(d)

Figure 6. F-measure scores for (a) b0, (b) b1, (c) b2, and (d) b3 classes

Table 3, tabulates the confusion matrixes of the CFS models. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. Furthermore, in Table 4 the true positive (TP) and false positive (FP) values for the

CFS models trained with C4.5, IB1, naive Bayes and Bayesian networks are tabulated.

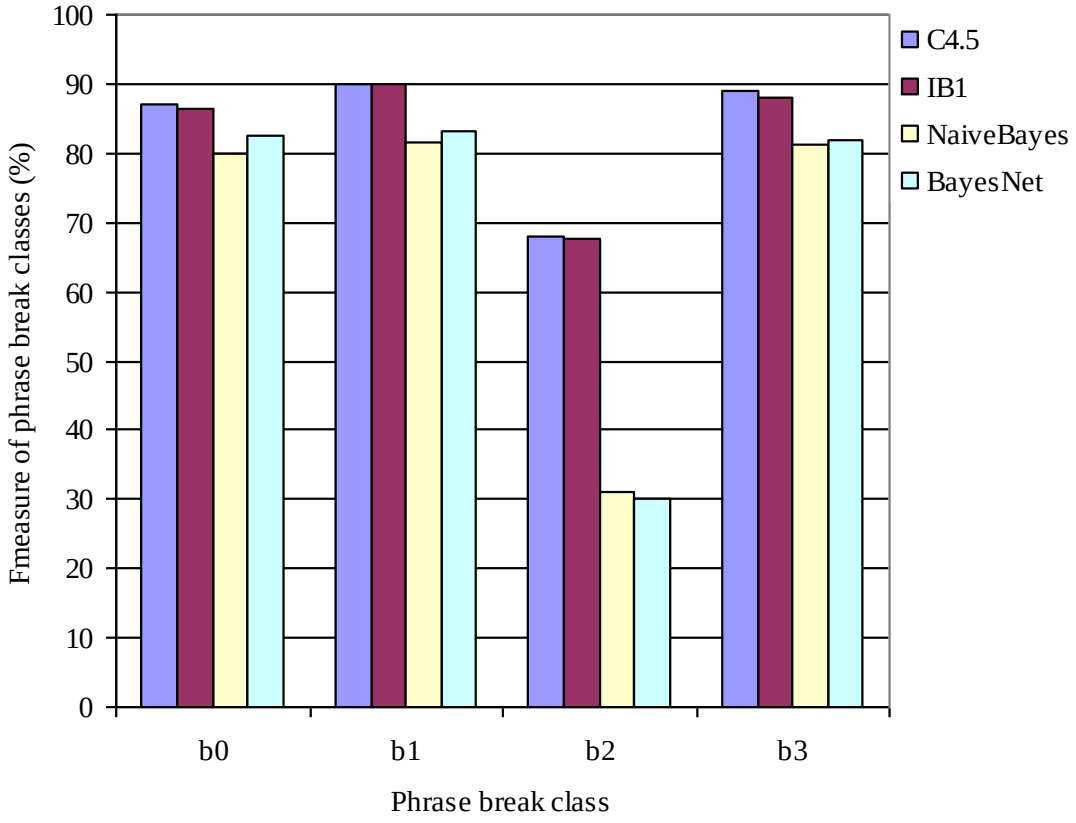


Figure 7. F-measure scores of CFS subset trained models

An interesting remark that can be extracted from Table 3 and Table 4 is that Bayesian methods confuse less the non-break categories with the breaks compared to C4.5 and IB1. In particular the FP scores of b2 and b3 are lower in the case of naive Bayes and Bayesian networks compared to C4.5 and IB1. Additionally, Table 3 shows that C4.5 tree inducer confuses less the non-break categories with the break categories compared to IB1. Table 4 clearly displays that IB1 has the lowest FP score for b1 class compared to all other approaches. In contrast, C4.5 showed the highest TP for this class.

Table 3. Confusion matrix of CFS subset trained models

C4.5	b0	b1	b2	b3
b0	1332	178	6	8
b1	160	4367	91	22
b2	32	438	822	77
b3	12	87	137	1395
NaiveB	b0	b1	b2	b3
b0	1336	186	2	0
b1	446	4099	87	8
b2	34	927	313	95
b3	8	200	238	1185
IB1	b0	b1	b2	b3
b0	1352	145	20	8
b1	202	4242	166	30
b2	36	335	878	121
b3	14	53	158	1405
BNet	b0	b1	b2	b3
b0	1330	194	0	0
b1	337	4252	48	4
b2	32	945	287	105
b3	6	204	214	1207

Table 4. TP and FP values for CFS subset trained models

C4.5	b0	b1	b2	b3
FP (%)	2.7	15.5	3	1.4
TP (%)	87.4	94.1	60.1	85.5
NaiveB	b0	b1	b2	b3
FP (%)	6.4	29	4.2	1.4
TP (%)	87.6	88.3	22.9	72.7
IB1	b0	b1	b2	b3
FP (%)	3.3	11.8	4.4	2.1
TP (%)	88.7	91.4	64.1	86.1
BNet	b0	b1	b2	b3
FP (%)	4.9	29.7	3.4	1.4
TP (%)	87.3	91.6	21	74

6.3. Practically Optimal Dataset Evaluation Results

Although total accuracy scores, for all machine learning schemes, attests the efficiency of IG ranking (and CFS filtering), there were cases where a particular feature although possessing a high IG rank (or selected in CFS), its application tends to lower the overall classification performance (mainly a result of the correlation between features). Furthermore, features with low IG did not contribute significantly to the overall performance of the prediction model (i.e. features from 46 to 70). For example, in the case of C4.5 models, Figure 6.d, addition of nn.gpos (which is in 14 position of the feature ranking table) seems to lower the classification performance from 84.1% to 82.1%.

For the selection or omission of features performance of all approaches from all the carried out experiments (IG datasets as well as CFS subset) was taken into account. This procedure led us to the construction of a “practically” optimal dataset that is consisted of the following features:

brk.pnct, word_out, POS, word.stress.strct, word.numsyls, ssyl.out, p.brk.pnct, n.position.type, chunk.neighb, n.chunk.dist, chunk.dist, p.last.ssyl.in.phrase, syl.in, , chunk, syl.codasize, pp.fc.POS, p.chunk.in, pp.last.ssyl.in.phrase. It worths mentioning that although certain phonological (syllabic) features were ranked in low positions by IG ranking or not selected from CFS filtering appeared to contribute (as shown in Figure 5 and Figure 6), thus included to the "practically" optimal dataset. Such features were syl.codasize, ssyl.out, syl.in as well as the contextual pp.last.ssyl.in.phrase.

Evaluation of the the "practically" optimal dataset was carried out with the WCL1 database and a limited domain prosodic database [32] that contains prosodic phenomena encountered in a museum guided tour. Both corpora were cross-checked for their annotation consistency [33].

Table 5 tabulates the total accuracy of C4.5, IB1, naive Bayes and Bayesian network models trained with the "practically" optimal feature set for both prosodic databases. It shows that C4.5 phrase prediction model performed better compared to all the other algorithms for both training domains followed by the Bayesian network model. Although *"practically"* optimal feature set was extracted empirically from experiments with the WCL1 database, limited domain models presented higher total accuracy prediction scores for all approaches; this can be explained since breaks are described by simpler "rules" due to the restrictions of the domain compared to the generic characteristics of WCL1 text corpus.

Table 5. Total accuracy of WCL1 and limited domain models

Domain	C4.5	IB1	Naïve Bayes	BayesNet
WCL1 (%)	88.77	78.91	77.94	82.46
Limited (%)	90.5	83.34	79.46	83

Figure 8 depicts the F-measure of each break class for WCL1 dataset. It is interesting to detail C4.5 performance regarding b2 class prediction; for this particular class it achieved an F-measure score of 75% while IB1, naive Bayes and Bayesian networks scored 50%, 42.3% and 56.9% respectively.

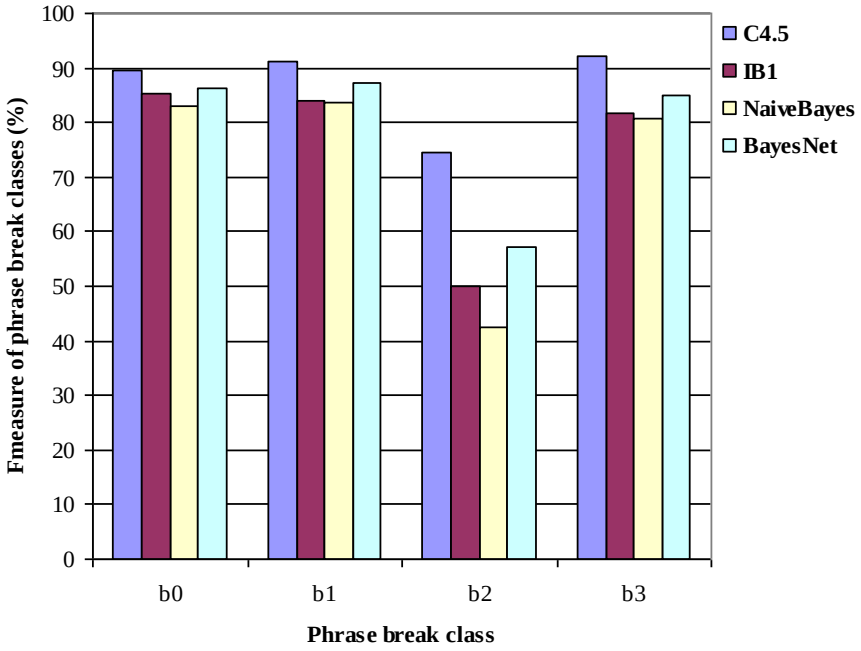


Figure 8. F-measure of WCL1 models with practically optimal feature set

Table 6 tabulates the confusion matrixes for each machine learning approach while in Table 7 the FP and TP scores of the phrase prediction models are presented.

Table 6. Confusion matrix of the WCL1 models

C4.5	b0	b1	b2	b3
b0	1358	155	11	0
b1	131	4359	123	27
b2	22	329	943	75
b3	4	61	89	1477
NaiveB	b0	b1	b2	b3
b0	1312	192	10	10
b1	283	4115	143	99
b2	36	757	470	107
b3	10	149	226	1246
IB1	b0	B1	b2	b3
b0	1318	176	30	0
b1	208	3919	428	85
b2	32	464	666	208
b3	6	121	174	1329
BNet	B0	B1	b2	b3
b0	1306	194	24	0
b1	178	4252	204	6
b2	14	551	735	69
b3	2	127	238	1264

Table 7. TP and FP values for the WCL1 models

C4.5	b0	b1	b2	b3
FP (%)	2.0	12	2.9	1.4
TP (%)	89.1	93.9	68.9	90.5
NaiveB	b0	b1	b2	b3
FP (%)	4.3	24.3	4.9	2.9
TP (%)	86.1	88.7	34.3	76.7
IB1	b0	b1	b2	b3
FP (%)	3.2	16.8	8.1	3.9
TP (%)	86.5	84.5	48.6	81.5
BNet	b0	b1	b2	b3
FP (%)	2.5	19.3	6.0	1.0
TP (%)	85.7	91.6	53.7	77.5

As shown in Table 7, C4.5 scored the lowest and highest scores of FP and TP respectively for all phrase break class compared to all the utilized machine learning approaches.

Figure 9 presents the F-measure results for phrase prediction obtained from C4.5, IB1, naive Bayes and Bayesian networks trained with the "practically" optimal set of features for the limited domain data. This figure

clearly proves the effectiveness of this feature set and in this case.

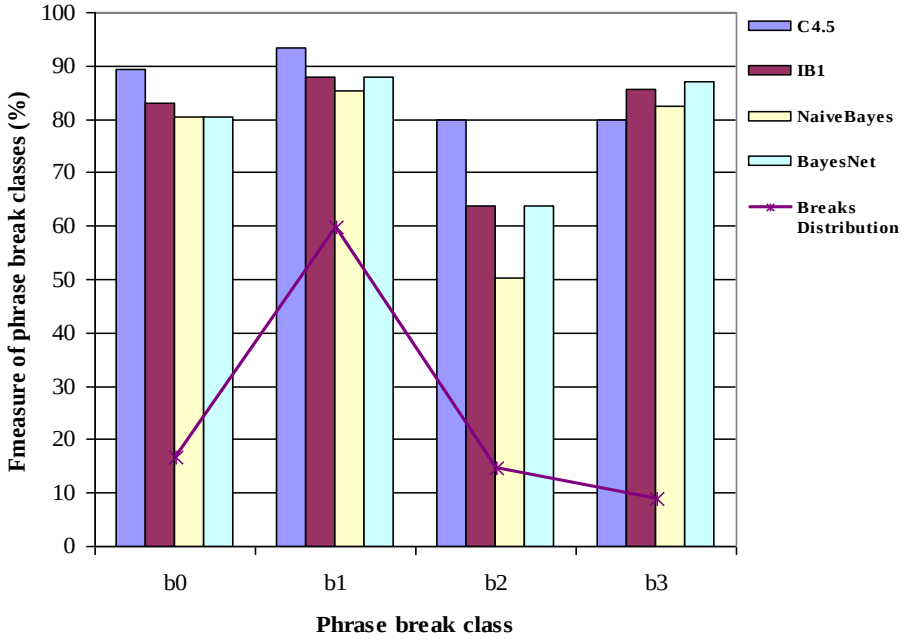


Figure 9. F-measure of limited domain models with practically optimal feature set

Table 8. Confusion matrix of the limited domain models

C4.5	b0	b1	b2	b3
b0	1060	127	3	0
b1	116	4053	125	3
b2	8	197	829	28
b3	0	16	59	559
NaiveB	b0	b1	b2	b3
b0	1003	151	29	7
b1	261	3690	315	31
b2	30	472	501	59
b3	6	30	84	514
IB1	b0	b1	b2	b3
b0	997	177	16	0
b1	201	3768	312	16
b2	15	297	686	64
b3	0	22	77	535
BNet	b0	b1	b2	b3
b0	948	204	38	0
b1	193	3797	300	7
b2	20	313	696	33
b3	3	27	83	521

The "practically" optimal dataset showed a comparable performance for both prosodic databases in all learning schemas. As regards the prediction of b1 and b2 categories, limited domain models achieved higher results than those scored by the WCL1 models. This is explained by the fact that WCL1 is composed of more complex prosodic events than that of the limited domain prosodic corpus.

Finally, in Table 8 and Table 9 the confusion matrix and the FP and TP scores of the limited domain models are tabulated. As in WCL1 datasets, and in the case of limited domain datasets the C4.5 model showed the lowest FP and the highest TP scores.

Table 9. TP and FP values for the limited domain models

C4.5	b0	B1	b2	b3
FP (%)	2.1	11.8	3.1	0.5
TP (%)	89.1	94.3	78.1	88.2
NaiveB	b0	b1	b2	b3
FP (%)	5	22.6	7	1.5
TP (%)	84.3	85.9	47.2	81.1
IB1	b0	b1	b2	b3
FP (%)	3.6	17.2	6.6	1.2
TP (%)	83.8	87.7	64.6	84.4
BNet	b0	b1	b2	b3
FP (%)	3.6	18.8	6.9	0.6
TP (%)	79.7	88.4	65.5	82.2

7. Conclusions

In this article, feature and algorithm evaluation was conducted for the task of intonational prosodic boundaries prediction for the Greek language. Initially, the utilized prosodic corpus was analyzed and textual, lexical, morphological and shallow syntactical features were extracted on word and syllabic level. Features contribution to the task was measured with the utilization of information gain and correlation based feature subset methods. From the feature ranking we constructed a total of 70 datasets while filtering method outputted an "optimal" subset of features. All datasets were applied to C4.5, IB1, naive Bayes and Bayesian network learning schemas. Taking into account the resulted total accuracy of all prediction models we were led to the construction of a "practically" optimal set of features. The effectiveness of "practically" optimal feature set was evaluated with WCL1 database as well as with a limited domain prosodic corpus.

Our plans for future work include the evaluation of the proposed phrase break models on the speech rendering procedure of our TtS with the utilization of acoustic tests (pitch analysis of the synthetic waveform) as well as perceptual tests with subjective listening tests. Furthermore we work upon the extension of WCL1 prosodic corpus with the addition of more annotated recording of the same and different speakers.

References

[1] Dutoit, T., An Introduction to Text-To-Speech Synthesis, Dordrecht, Kluwer Academic Publishers, 1997.

[2] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirsberg, J., "ToBI: A standard for labelling Eng¬lish prosody", Proceedings of the International Conference on Spoken Language Processing, Alberta, October 13-16, 1992, vol. 2, p. 867-870.

[3] Arvaniti, A., and Baltazani, M., "Greek ToBI: A System For The Annotation Of Greek Speech Corpora", Proceedings of Second International Conference on Language Resources and Evaluation, Athens, 31 May-2 June, vol. 2, 2002, p. 555-562.

[4] Bolinger, D., Intonation and its Uses: Melody in Grammar and Discourse, Stanford, Stanford University Press, 1989.

[5] Taylor, P., Black, A. W., "Assigning Phrase Breaks from Part-of-Speech Sequences", Journal of Computer Speech and Language, vol. 12, 1998, p. 99-117.

[6] Anderson, Stephen R. (1995). ``Rules and Constraints in Describing the Morphology of Phrases." Proceedings of the Chicago Linguistic Society, vol. 31 (Parasession volume on Clitics), pp. 15-31.

[7] Prieto, P., Hirschberg, J., "Training Intonational Phrasing Rules Automati¬cally for English and Spanish text-to-speech", Journal of Speech Communication, vol. 18, issue 3, 1996, p. 281-290.

[8] Bachenko, J., and Fitzpatrick, E., "A computational grammar of discourse-neutral prosodic phrasing in English", Journal of Computational Linguistics, vol. 16, Issue 3, 1990, p. 155-170.

[9] Ostendorf, M., Veilleux, N., M., "A hierarchical stochastic model for automatic prediction of prosodic boundary location", Journal of Computational Linguistics, vol. 20, issue 1, 1989, p. 26-53.

[10] Riley, M., "Tree-based modelling of segmental duration", in Talking Machines: Theories, Models, and Designs, G. Bailly and C. Benoit, Eds. Elsevier Science Publishers, 1992, pp. 265-273.

[11] Muller, A., F., Zimmermann, H., G., and Neuneier, R.,: 1996, "Robust Generation of Symbolic Prosody by a Neural Classifier Based on Autoassociators", Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Atlanta, GA, May 7-10, 1996, p. 1285-1288.

[12] Fordyce, C., S., Osterdorf, M., "Prosody Prediction for Speech Synthe¬sis Using Transformational Rule-Based Learning", Proceedings of International Conference on Spoken Language Processing, Sydney, 30 November-4 December 1998, p.682-685.

[13] Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., "Hidden Markov models based on multi-space probability distribution for pitch pattern modelling". Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Pheonix, March 15-19, 1999, p. 229-232.

[14] Busser, Bertjan, Daelemans, Walter, Bosch, Antal van den (2001): "Predicting phrase breaks with memory-based learning", In SSW4-2001, paper 125.

[15] Zervas, P., Maragoudakis, M., Fakotakis, N., and Kokkinakis, G., "Bayesian Induction of intonational phrase breaks", Proceedings of Eurospeech, Geneva, September 1-4, 2003, p. 113-116.

[16] Quinlan, J., R., C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Francisco, 1993.

[17] Aha, D., Kibler, D., Albert M., "Instance-based learning algorithms", Machine Learning, vol. 6, 1991, p. 37-66.

[18] Domingos, P., Pazzani, M., "Beyond independence: Conditions for the optimality of the simple bayesian classifier", Proceedings of the Thirteenth International Conference on Machine Learning, Bari, July 3-6, 1996, p. 105-112.

[19] Cowell, R., Dawid, A., P., Lauritzen, S., L., Spiegelhalter, Probabilistic networks and expert systems, Springer, 1999.

[20] Cormen, T., Leiserson, C., and Rivest, R., :1990, Introduction to Algorithms, MIT Press, Chap. 16, Greedy Algorithms

[21] Price P. J., Ostendorf M., Shattuck, Hufnagel S., Fong., The use of prosody in syntactic disambiguation, J. Acoust. Soc. Am. Volume 90, Issue 6, pp. 2956-2970, 1991.

[22] Hirschberg, J., (1993) "Pitch accent in context: predicting intonational prominence from text", Artificial Intelligence 63, pp. 429-432.

[23] Prevost, S., (1995) "A semantics of contrast and information structure for specifying intonation in spoken language generation", PH.D. Thesis, University of Pennsylvania, 1995.

[24] Sgarbas, K., Fakotakis, N., Kokkinakis, G., "A morphological description of MG using the two-level model", Proceedings of the 19th Annual Workshop, Division of Linguistics, Thesaloniki, April 23-25, 1999, p.419-433.

[25] Stamatatos, E., Fakotakis, N., Kokkinakis, G., "A Practical Chunker for Unrestricted Text", Proceedings of the Second International Conference on Natural Language Processing, Patras, June 2-4, 2000, p. 139-150.

[26] Fujisaki, H., Nagashima, S., "A model for the synthesis of pitch con¬tours of connected speech" Annual Report of the Engineering Research Institute, University of Tokyo, 1969, pp. 53-60,

[27] Taylor, P., "The rise/fall/connection model of intonation", Journal of Speech Communication vol. 15, 1995, pp. 169-186,

[28] Veronis J., Di Cristo Ph., Courtois F., Chaumette C., "A stochastic model of intonation for text-to-speech synthesis", Journal of Speech Communication vol. 26, 1998, pp. 233-244,

[29] Taylor, P., "Analysis and synthesis of intonation using the Tilt model", Journal of the Acoustical Society of America , vol. 107, issue 3, 2000, pp. 1697-1714,

[30] Hart, J.,'t, Collier, R., "Integrating different levels of intonation analysis", Journal of Phonetics, vol. 3, 1975, pp. 235-255.

[31] Alessandro, C., d', Mertens, P., "Automatic pitch contour stylization using a model of tonal perception", Computer Speech and Language, vol. 9, issue 3, 1995, pp. 257-288.

[32] Xydas G., Spiliotopoulos D., Kouroupetroglou G.: "Modelling Improved Prosody Generation from High-Level Linguistically Annotated Corpora", IEICE Transactions of Information and Systems, 2005, p. 510-518.

[33] Zervas, P., Xydas, G., Fakotakis, N., Kokkinakis, G., Kouroupetroglou, G., "Evaluation of Corpus Based Tone Prediction in Mismatched Environments", Proceedings of 8th International Conference on Spoken Language Processing, Jeju, October 4-8, 2004, p. 761-764.

[34] Witten, I., H., Frank, E., Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005

[35] Breiman, L., Friedman, J., H., Olshen, R., A., Stone C., J., Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.

[36] Quinlan, J., R., "Induction of decision trees", Journal of Machine Learning, vol. 1, 1986, p. 81-106.

[37] Palmer, D., Hearst, M., "Adaptive Multilingual Sentence Boundary Disambiguation", Journal of Computational Linguistics, vol. 23, issue 2, 1997, p. 241-267.

[38] Brill, E., "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging", Journal of Computational Linguistics, vol. 21, issue 4, 1995, p. 543-565.

[39] Magerman, D., "Statistical Decision-Tree Models for Parsing", Proceedings of Meeting of the Association for Computational Linguistics, MIT, Cambridge, Massachusetts, 26-30 June, USA, p. 276-283.

[40] Black, A., Taylor, P., "Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input", Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, 18-22 September, 1994, vol. 2, p. 715-718.

[41] Lee, S., Oh, Y., "Tree-based modeling of prosodic phrasing and seg¬mental duration for Korean TTS systems", Journal of Speech Communication, vol. 28, issue 4, 1999, p. 283-300.

[42] Mitchell T., Machine Learning, Mc Graw-Hill, 1997.

[43] Kohavi, R., John, G., H., "The Wrapper Approach", in Feature Selection for Knowledge Discovery and Data Mining, H. Liu & H. Motoda (eds.), Kluwer Academic Publishers, 1998, p. 33-50.

[44] Blum, A., Langley, P., "Selection of relevant features and examples in machine learning", Artificial Intelligence, vol. 97, no. 1-2, 1997, p. 245-271.

[45] Hall, M., Smith, L., A., "Practical feature subset selection for Machine Learning", Proceedings of the Australian Computer Science Conference, February, 1996.

[46] Weiss, S., M., Kulikowski, C., A., Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems, Morgan Kaufmann, San Mateo, 1991.