Dealing with Imbalanced Data using Bayesian Techniques

Manolis Maragoudakis[§], Katia Kermanidis[§], Aristogiannis Garbis[‡] and Nikos Fakotakis[§]

[§]Artificial Intelligence Group
University of Patras
26500 Rion, Patras
{mmarag,kerman,fakotaki}@wcl.ee.upatras.gr

[‡]Department Of Applied Informatics In Management & Finance
Technological Education Institute of Messolonghi
Nea ktiria 30200, Messolonghi
agarbis@teimes.gr

Abstract

For the present work, we deal with the significant problem of high imbalance in data in binary or multi-class classification problems. We study two different linguistic applications. The former determines whether a syntactic construction (environment) co-occurs with a verb in a natural text corpus consists a subcategorization frame of the verb or not. The latter is called Name Entity Recognition (NER) and it concerns determining whether a noun belongs to a specific Name Entity class. Regarding the subcategorization domain, each environment is encoded as a vector of heterogeneous attributes, where a very high imbalance between positive and negative examples is observed (an imbalance ratio of approximately 1:80). In the NER application, the imbalance between a name entity class and the negative class is even greater (1:120). In order to confront the plethora of negative instances, we suggest a search tactic during training phase that employs Tomek links for reducing unnecessary negative examples from the training set. Regarding the classification mechanism, we argue that Bayesian networks are well suited and we propose a novel network structure which efficiently handles heterogeneous attributes without discretization and is more classification-oriented. Comparing the experimental results with those of other known machine learning algorithms, our methodology performs significantly better in detecting examples of the rare class.

1. Introduction

In the field of Natural Language Processing, research has been focusing on constructing systems that are able to treat large, naturally occurring texts. However, efficient handling of such applications requires rich lexical knowledge sources. Due to the fact that manual construction of lexicons containing linguistic information is laborious and time consuming, the recent research focus is the automation of the process, in terms of mining such information from available textual corpora.

The automatic acquisition of syntactic verb subcategorization frames (SF), i.e. the various syntactic entities a certain verb can be combined with in order to form a grammatical verb phrase and express its semantic arguments, is an interesting example of lexical acquisition. Dictionaries containing subcategorization information are essential especially for tasks like wide-coverage parsing, grammar development and text mining of higher-level (e.g. semantic) information.

Noise occurring in the input data due to ungrammaticalities of natural language, rarity of patterns of correct subcategorization information in the input text and finally errors that arise when processing the data (generating frame candidates and then selecting a valid set of frames for the lexicon) constitute challenging problems for the particular task.

Most of the previous approaches to frame acquisition have made use of sophisticated linguistic resources and pre-processing tools (syntactic treebanks, wide coverage parsers etc.). As, for the majority of languages (including Modern Greek), such resources are not yet available, acquiring the necessary information by making use of as limited linguistic resources as possible appears to be very challenging.

Previous work on the automatic acquisition of verb subcategorization information has focused mainly on statistical filtering methods applied to generated putative subcategorization frames (Brent, 1993; Briscoe and Carroll, 1997; Manning, 1993; Korhonen et al., 2000). Hypothesized frames are generated through methodologies varying in the extent of pre-processing of the corpus data used as input to the learning process, the number of frames learned, whether the set of frames is known beforehand or not.

Machine learning techniques have also been used for tasks closely related to the one at hand. Buchholz (1998) uses Memory-based learning on the parsed version of the Wall Street Journal to distinguish between complements and adjuncts, information fundamental for the automatic acquisition of subcategorization lexicons. Carroll and Rooth (1998) use a probabilistic headed context-free grammar (PCFG), based on chunk and phrase structures, complementation rules and n-gram rules and then apply the Expectation-Maximization (EM) algorithm iteratively to first train the grammar, and then yield probability estimates for verb and frame combinations. Miyata et al. (1997) propose a method for learning a Bayesian Network model of verbal subcategorization preference on Japanese in order to extract semantic collocational knowledge of verbs. Niyogi (2002) uses Bayesian analysis to learn the syntactic and semantic features that determine clusters of verbs that pattern together in the same constructions and describes in detail the theoretical modeling process of the features of a novel verb using limited training observations.

Name Entity Recognition (NER) is one of the most important information extraction tasks. By *Name Entity*, we denote the phrases that reside in textual corpora which can be used as names. Name Entities are actually noun phrases, however, not all noun phrases are Name Entities. Therefore, the role of NER is to recognize the noun phrases that are actually Name Entities and then categorize them into different partitions. 3 major classes have been identified i.e. person names, location names and organization names. The instances that belong to those classes are dominated by the instances that are not recognized as Name Entities (Japckowicz, 2000). (i.e. the negative class). Therefore, a methodology for effective coping with the high imbalance of data is required.

NER is mostly based to two linguistic sources. The former is a lexicon of known names, also called as "gazetteer" and the latter is a grammar that contains rules

for the identification of names (regardless of their existence within the lexicon). The grammar rules can be either hand-coded (i.e. they have been provided by domain experts) or generated by a machine learning algorithm (Karkaletsis et al., 1999).

We consider Bayesian networks theory in order to construct a classification framework. Furthermore, we describe a novel Bayesian network that is capable of coping with domains of discrete and continuous attributes without having to perform discretization, a process that often deteriorates the classification performance. A second novelty of this work is that introduction of a certain strategy for removing unnecessary negative training examples. We describe the whole process of identifying those examples and provide a methodology for obtaining a new training set, which causes a machine learning algorithm to generalize better.

2. Forming Input Data

2.1. Subcategorization Frames

The ILSP/ELEFTHEROTYPIA (Hatzigeorgiu et al., 2000) and ESPRIT 860 (Partners of ESPRIT-291/860, 1986) Corpora (a total of approximately 300,000 words) were used as input. Both corpora are balanced in genre and domain and manually annotated with complete morphological information. Further (phrase structure) information is obtained automatically by a phrase analyzer (chunker), described in detail in (Stamatatos et al., 2000), that detects noun, verb, prepositional and adverbial phrases. The chunker is based on a small keyword lexicon containing some 450 keywords (articles, pronouns, etc.) and a suffix lexicon of 300 of the most common word suffixes in MG. The head-word of every noun phrase is identified next, based on a set of empirical rules, and the phrase inherits its grammatical properties. As mentioned previously, phrases are non-overlapping. Verb complements are not included within the verb phrase, trailing adjuncts constitute a separate phrase, nominal modifiers in the genitive case are included within the noun phrase they modify, simple coordinate structures (parts of a noun phrase, for example, conjoined with a coordinating conjunction) build one phrase.

2.1.1. Feature Selection

As no other linguistic resources were utilized for the present work, the environment of the verb was determined empirically. After a number of experiments, concerning the window size of the environment, were carried out, a window of size (-2+3) was chosen, i.e. two phrases preceding and three phrases following the verb. Every verb-environment pair has been modeled via a set of nominal and numerical features that were extracted empirically. The nominal features contain information regarding the linguistic properties of the environment that affect the task at hand. The numerical features contain statistical information concerning the verb-environment co-occurrence in the data. In more detail, the nominal features are:

- the lemma of the verb
- a feature indicating the type of the verb
- a Boolean feature indicating whether the environment contains a prepositional phrase that is a priori known to never constitute a verb frame.
- a Boolean feature indicating whether the environment contains a prepositional phrase.
- a Boolean feature indicating whether the environment contains a punctuation mark, a symbol, foreign words etc.
- a Boolean feature indicating whether the environment contains a conjunction that introduces a nominal secondary clause.
- a Boolean feature indicating whether the environment contains a conjunction that introduces an adverbial secondary clause.
- a Boolean feature indicating whether the environment contains a nominal constituent (noun, adjective, numeral, personal pronoun) in the nominative case.
- a Boolean feature indicating whether the environment contains a nominal constituent (noun, adjective, numeral, personal pronoun) in the accusative case.
- a Boolean feature indicating whether the environment contains a coordinating conjunction.
- a feature indicating whether the environment contains a pronoun and its type (relative, interrogative etc).
- a feature indicating whether the environment contains an adverb and its type (modal, temporal etc).

The numerical features are:

- the size of the environment
- $p_1 = k_1 / n_1$, where k_1 is the count of co-occurrences of verb v with environment e and n_1 is the total number of occurrences of verb v in the data.
- $p_2 = k_2 / n_2$, where k_2 is the count of co-occurrences of every other verb except for v with environment e and n_2 is the total number of occurrences of every other verb except for v in the data.
- LLR=[logL(p1,k1,n)+logL(p2,k2,n)-logL(p,k1,n)-logL(p,k2,n)] where logL(a,b,c)=clog(a)+(b-c)log(1-a).

2.2. Name Entity Recognition

We used the WCL NER database (Tasikas, 2002). The framework of the database was the ILSP/ELEFTHEROTYPIA corpus with minimal preprocessing. More specifically, the titles of the articles, the named of the editors the source of the article, etc. had been manually removed resulting in a set of 131,545 lexical token, with embedded morphological information. Manual annotation of the lexical units with Name Entity classes (Name, Organization, and Location) was then followed, abiding with the MUC-7 contest directives.

2.2.1. Feature Selection

A window size of [-1,+1] was found to provide the best results. The database contained the following features:

- part-of-speech: in MG it takes 13 different values, including punctuation marks and foreign words
- case: it takes 6 values (nominative, genitive, accusative, vocative, dative and a null one for the indeclinable words)
- number: with 3 different values (singular, plurar and a null one for tokens without a number)

- tokenType: this variable identifies the writing type of a lexical unit. It has a total of 11 different values.
- period: it refersto the lemma of the lexical unit ".". It is a binary variable.
- abbreviations: contains the lemma of the abbreviations of the corpus.

3. Mixed Gaussian Bayesian Augmented Naïve Bayes

As for a classification algorithm, Bayesian networks (Heckerman, 1996; Cheng & Greiner, 1999) were chosen. However, Friedman and Goldszmidt (1996) have noticed that a general, unrestricted Bayesian network may not be good classifier. They justify this by observing that learning a Bayesian network structure is unsupervised in the sense that no distinction is made among the class node and the other attribute nodes. In other words, since the class node is not explicitly stated, the structure – or some part of it – may not be relevant for classification, if it is outside the Markov blanket of the class node. Pearl (1988) defined the Markov blanket of a node x as the union of x's direct parents, x's direct children and all direct parents of x's children. The semantic notion of the Markov blanket is that x is unaffected by nodes outside of its Markov blanket Madden, 2002). In order to alleviate this problem, we propose a new Bayesian network structure that can cope with heterogeneous attributes and is more classification oriented, named as "Mixed Gaussian Bayesian Augmented Naïve Bayes-(mG-BAN)"

We begin by assuming a set of mixed attributes A is partitioned as $A = \Delta \cup \Gamma$ into discrete (Δ) and continuous (Γ) variables. Furthermore, we suppose that the joint distribution of the continuous variables follows a multivariate Gaussian distribution such as:

$$f(\Gamma \mid \Delta = i) = N_{|\Gamma|}(\vec{\mu}_i, \Sigma_i), \quad \text{where} \quad N_{|\Gamma|}(\vec{\mu}_i, \Sigma_i)$$

denotes the multivariate Gaussian distribution with mean $\vec{\mu}_i$ and covariance matrix Σ_i and $|\Gamma|$ is the cardinality of *Γ*. Note that the discrete nodes cannot have continuous parents in this model. This is the most general case where exact inference algorithms are known (Murphy, 1998). Moreover, if a continuous node *Γ* has a discrete parent *Δ*, it has a different mean and covariance matrix for every distinct value of *Δ*.

The proposed mixed Gaussian Bayesian Augmented Naïve Bayes (mG-BAN) construction algorithm is composed of the following steps:

Again, we assume fully observable attribute values. Given a training set *D*, consisted of discrete (*Δ*) and continuous (*Γ*) attributes, and a nominal class (*C*):

- Use the Cooper and Herskovits (1992) Bayesian scoring function (the MDL scoring metric has proved to be asymptotically equivalent with it as the number of instances grows, so there is actually no difference in the metric that one might choose) and return the most probable network structure B_Δ (this task's complexity is $O(\delta^2)$, where δ is the number of discrete attributes).
- Add *C* as a parent of every discrete attribute Δ_i.
- Estimate the parameters of that structure, using the empirical conditional frequencies from the data (Cooper & Herskovits, 1992).
- Add *C* as a parent of every continuous attribute Γ_i.
- For every class label c of the class node C, estimate the mean and variance of each continuous attribute Γ_i. Update the existing conditional probability tables with the extracted means and variances. For a binary class domain, this usually takes $O(2\gamma)$, with γ to denote the number of continuous variables.
- Output the mG-BAN for the given training set.

Recall that learning a general, unrestricted Bayesian network from hybrid data requires $O(\delta^2 + \gamma^2)$ time whereas in our proposed structure this task is far quicker ($O(\delta^2 + 2\gamma) \approx O(\delta^2)$).Moreover, we overcome the potential difficulties that discretization poses to the classification task by incorporating multivariate Gaussian distributions for each class value. This step is theoretically sound, since it allows for better capturing of the real continuous data distributions.

4. Highly imbalanced instances

In order to portray the negative influence in terms of classification accuracy that abundant negative instances pose, consider the artificial data set of Figure 1. Points that are marked with (+) denote positive instances while those marked with (-) denote the negative examples. The dashed line is often referred to as borderline, which serves the role of a decision surface over the set of examples.

Assuming a Bayesian-based classifier (Narayanan & Jurafsky, 1998), it needs – by principle – to calculate the prior probabilities of the positive and negative class P(+) and P(-) respectively. Denote also by pdf(+)(x) and pdf(-)(x) the probability density functions for a given point x, for each class respectively. Without taking any misclassification costs into consideration, the algorithm would classify an instance y as positive if and only if: $P(+)*pdf_{(+)}(x) > P(-)*pdf_{(-)}(x)$

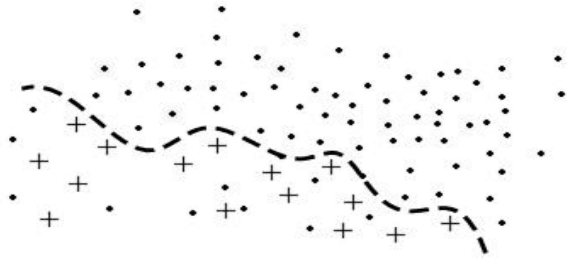

Figure 1. Visualization of instances of the artificial data.

The presence of plentiful negative examples in the training set would result in P(-)>>P(+) thus the above inequality could hardly ever be satisfied, even if $pdf_{(+)}(x) > pdf_{(-)}(x)$. A stereotypical approach to deal with this problem would be to assign very large costs to misclassifying positive examples (cost(+)) and associate a very small cost (cost(-)) with the negative examples. In that case, the Bayesian classifier would classify y as positive whenever:

$$P(+)*pdf_{(+)}(x)*cost(+) > P(-)*pdf_{(-)}(x)*cost(-)$$

Nevertheless, even in that case, the classifier would have problems estimating a smooth density function of the positive class. Besides, assignment of empirical costs is not clearly determined prior to training.

4.1. Abundant Examples

The primary motivation of the proposed methodology is to end up with a representative training set where each class distribution will not suffer from disproportions. Thus, the misclassification problems described in the previous paragraphs would be alleviated. In order to obtain such a representative data set, we applied a selection technique that was first introduced in 1976 by Tomek and was later applied in various machine learning research studies (e.g. Aha et al., 1991; Skalak 1994; Lewis & Gale 1994; Floyd & Warmuth 1995; Kubat & Matwin 1997; Suzuki, 1993). Despite the fact that most of these contributions were attempting to reduce training size, it can be safely applied to our domain, with the only requirement that all positive instances should be maintained in the training set, since they are too rare thus too precious to be discarded. Regarding the negative instances, they can be characterized into four different groups (Figure 2):

- Noisy: It contains examples that are situated within a cluster of examples of the opposite class.
- Borderline: It contains examples that are close to the boundary region between two classes.
- Redundant: It contains examples that can be already described by other examples of the same class.
- Safe: It contains examples that are crucial for determining the class, thus needed for the training stage.

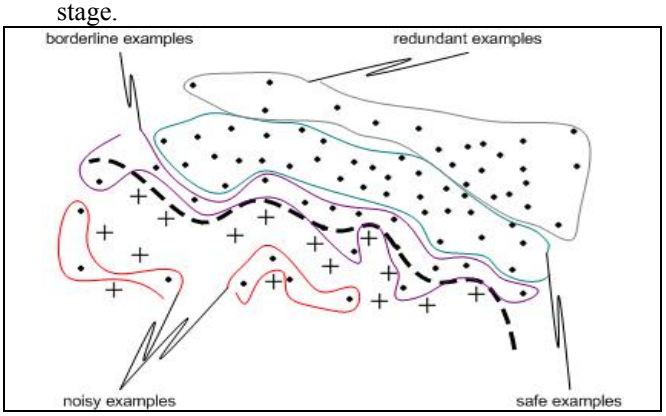

Figure 2. Characterising negative instances.

Our idea is focused on preserving all positive examples plus only the safe negative examples by removing the redundant, borderline and particularly the noisy instances. In order to perform this task, we apply the notion of Tomek links (Tomek, 1976). A Tomek link is a pair of two examples x and y of different classes, where no other example z exists such that: $\delta(x,z) < \delta(x,y)$ or $\delta(y,z) < \delta(x,y)$, where δ(x,y) denotes the distance between vectors x and y.

The exact process of the proposed algorithm is:

1. Let T be the original training set, where $\|T_{(-)}\| >> \|T_{(+)}\|$, meaning that the size of the negative examples outnumbers that of the positive examples.

2. Construct a data set C, containing all positive instances plus one randomly selected negative instance.

3. Classify T with the mG-BAN classifier using the training examples of C and move all misclassified items to C. C is still consistent with T only smaller. (Note that one may use a classifier of his/her preference).

4. Remove each negative example that is found to be participating in a Tomek link with every positive example, out of $h\%$ percentage of the population of positive examples in a neighboring cluster of such examples.

5. The resulting set T_{opt} is used for classification instead of T.

5. Evaluation and Experimental Results

In order to evaluate the behaviour of the subcategorization frame module, we extracted all sentences containing one of the following thirty verbs: *αισθάνομαι* (feel), *φοβάμαι* (be afraid of), *μιλώ* (speak), *υπόσχομαι* (promise), *ξέρω* (know), *φαίνομαι* (seem), *μοιάζω* (resemble), *πιστεύω* (believe), *σκέπτομαι* (think), *βοηθώ* (help), *μαθαίνω* (learn), *θυμάμαι* (remember), *λέγω* (say), *δηλώνω* (announce), *μένω* (stay), *αποφασίζω* (decide), *κάνω* (do), *βλέπω* (see), *ακούω* (hear), *δείχνω* (show), *προτείνω* (suggest), *θεωρώ* (consider). The verbs were chosen randomly; provided that they appeared a sufficient number of times in the corpus and that they presented a variety in syntactic arguments. The sentences were manually tagged (arguments distinguished from adjuncts) by specialists. The final data set proved to be highly imbalanced: positive instances (denoting a valid frame) and negative instances appear in the data set with a ratio of 1:80 respectively (About 450 positive examples and 38.000 negative examples).

Regarding the NER application, the ratio of positive instances (one of the three main classes) and the negative ones (the null class) was 1:120 (About 3500 Name Entities in 415,000 candidate noun phrases). For both applications we calculated per class precision, per class recall and F-measure as a harmonic mean of the two. Accuracy in some domains, such as the one at hand, is not actually a good metric due to the fact that a classifier may reach high accuracy by simply always predicting the negative class.

A set of well-known machine learning techniques have constituted the benchmark to which our results have been compared: Naïve Bayes, Decision Trees (C4.5), and k-NN instance-based learning. No pruning was performed when using C4.5 so as to make sure rare instances are not overlooked. Cross validation was performed with k-NN in order to determine the best k. AdaBoost (Schapire, 1999) was also applied as a meta-learning boosting algorithm to cope with the numerous negative instances and was compared to the variation of our approach to boosting that utilizes a variation of the algorithm by (Kubat & Matwin, 1997) based on Tomek links. The results were obtained using 10-fold cross validation, where the ration of positive and negative examples was kept equal for all ten tests. Figures 3 and 4 portray the performance for the SF application while figures 5 and 6 illustrate the performance metrics for the NER domain. In the latter, the positive class has grouped the three main classes (name, location and organization).

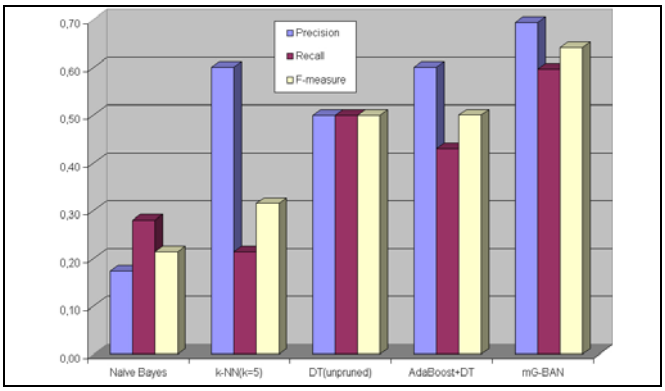

Figure 3. Performance of the SF positive class.

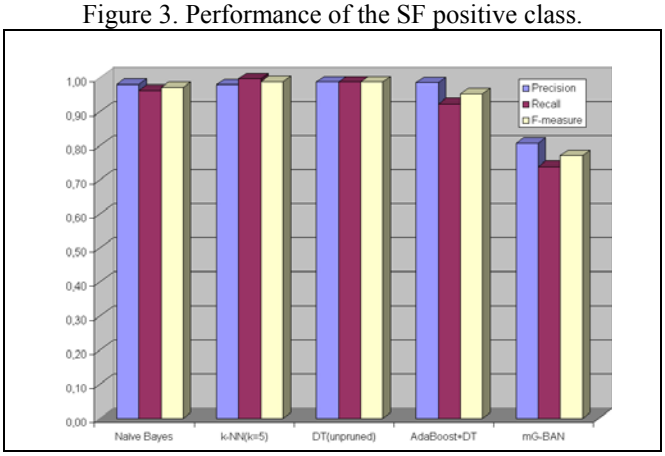

Figure 4. Performance of the SF negative class.

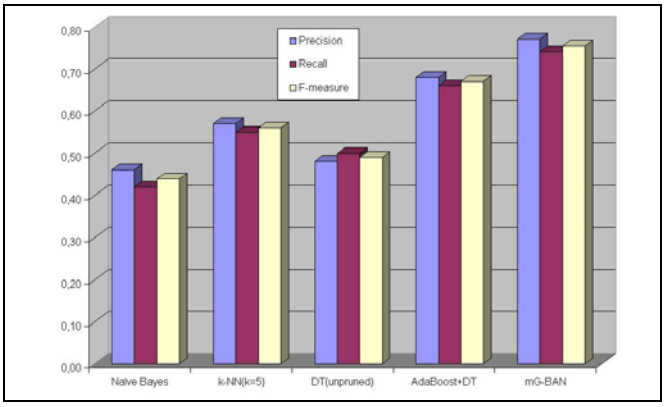

Figure 5. Performance of the NER positive class.

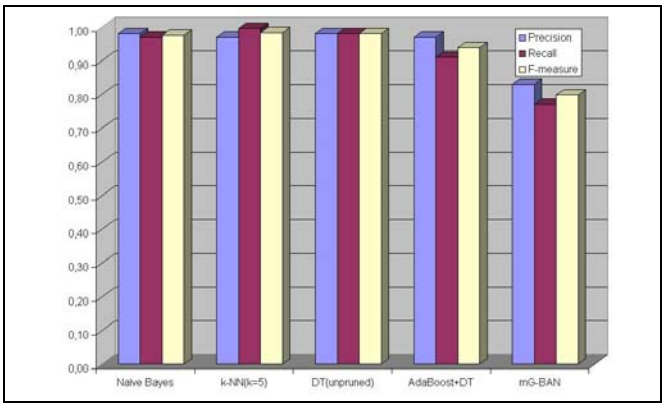

Figure 6. Performance of the NER negative class.

As can be observed in the above diagrams, imbalance in the data affects accuracy of the positive class significantly when applying traditional learning algorithms. C4.5 with no pruning performs better than the rest of the traditional methods due to the fact that rare instances are not disregarded. AdaBoost copes better with the imbalance problem by improving the performance in labeling the positive instances up to 10%. We additionally observe that Naïve Bayes performs very well on the negative class whereas regarding the positive examples, it proves to be the worst. This can be justified by the argument that Naïve Bayes uses probability distributions of the sparse positive instances (John & Langley, 1995).

On the other hand, the mixed Gaussian BAN structure appears to relax the above assumptions and furthermore, it better captures the real distributions of the continuous variables, while in Naïve Bayes discretization did not seem to have finely quantized the continuous values.

By removing from the data set the negative instances that participate in Tomek links, we disregard the abundant negative examples in the borderline area and thus reduce the bias of the classifier towards the negative class. This results to a more balanced data set and the increase in recall and precision for labeling positive instances is significant. The effect of the reduction of the training size is portrayed in terms of lower performance of the negative class. However, this error rate is much smaller in relation to the gain of correctly identifying the positive examples.

6. Conclusion

In this paper, we have proposed a new methodology for creating Bayesian network structures that perform well on classification tasks. This new structure, which we call mixed Gaussian Bayesian Augmented Naïve Bayes (mG-BAN), is capable of efficiently handle domains of discrete and continuous variables without having to perform discretization. Furthermore, the complexity of this task is significantly faster than that of learning a general Bayesian network from data. We applied mG-BAN to a text mining task where correctly classifying the positive instances is deteriorated since the imbalance of positive and negative instances is very high. As a method for dealing with this problem, we have applied a strategy that has not been utilized in linguistic domains before. More respectively, under-sampling of the abundant negative examples has been carried out, based on Tomek links. Furthermore, we claim that the proposed algorithms can also be applied to other domains that present similar behavior to the one we examined, such as spotting of credit card fraud, identifying oil-spills from satellite images, or the thyroid disease domain (from the UCI repository, Murphy & Aha, 1993). Our future goal is to apply the mG-BAN algorithm to such applications and publish a software version of it, in order to be available to other researchers.

7. References

Aha, D., Kibler, D. & Albert, M.K. (1991). Instance based learning algorithms. Machine Learning, 6, pp. 37-66.

Brent, M. (1993). From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. Computational Linguistics, 19, pp. 243-262.

Briscoe, T. & Carroll, J. (1997). Automatic Extraction of Subcategorization from Corpora. Proceedings of the 5th

ANLP Conference, ACL,. Washington D.C, pp. 356-363.

Buchholz, S. (1998). Distinguishing Complements from Adjuncts using Memory-Based Learning. Proceedings of the Workshop on Automated Acquisition of Syntax and Parsing, ESSLLI-98, Saarbruecken, Germany, pp. 41-48.

Carroll, G. & Rooth, M. (1998). Valence Induction with a Head-Lexicalized PCFG. Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP 3), Granada, Spain, pp. 36-45.

Cheng, J. & Greiner, R. (2001). Learning Bayesian Belief Network Classifiers: Algorithms and System. Proceedings of the Canadian Conference on Artificial Intelligence, Ottawa, Canada.

Cooper, J. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9, pp. 309-347.

Floyd, S. & Warmuth, M. (1995). Sample Compression, Learnability and the Vapnik-Chervonenkis Dimension. Machine Learning, 21, pp. 269-304.

Friedman, N. & Goldszmidt, M. (1996). Discretizing continuous attributes while learning Bayesian networks. In L. Saitta, (eds), Machine Learning: Proceedings of the Thirteenth International Conference, Morgan Kaufmann Publishers, pp. 157-165.

Friedman, N. & Goldszmidt, M. (1996). Discretizing Continuous Attributes while Learning Bayesian Networks, Proceedings of the 13th International Conference on Machine Learning, pp. 157-165.

Hatzigeorgiu, N., Gavriilidou M., Piperidis S., Carayannis G., Papakostopoulou, A. Spiliotopoulou, A. et al. (2000). Design and Implementation of the online ILSP Greek Corpus. Proceedings of LREC 2000, Athens, Greece, pp. 1737-1742.

Heckerman, D. (1996). A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmond, Washington.

Japkowicz, N. (2000). The Class imbalance problem: Significance and strategies. In proceedings of the 2000 international conference on Artificial intelligence, Special Track on Inductive Learning, Las Vegas, Nevada.

John, G., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338-345.

John, G., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338-345.

Karkaletsis, V., Paliouras, G., Petasis, G., Manousopoulou, N. and Spyropoulos, C.D. (1999). Name Entity Recognition from Greek and English Texts, Journal of Intelligent and Robotic Systems, 26, pp. 123-135.

Korhonen, A., Gorrell, G. & McCarthy D. (2000). Statistical filtering and subcategorization frame acquisition. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Hong Kong, pp. 199-205.

Kubat, M. & Matwin, S. (1997). Addressing the curse of imbalanced training sets. Proceedings of the International Conference on Machine Learning, pp. 179- 186.

Lewis, D. & Gale, W. (1994). Training Text Classifiers by Uncertainty Sampling. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Madden, M.G. (2002). Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm. Technical Report No. NUIG-IT-011002, Department of Information Technology, National University of Ireland, Galway.

Manning, C. (1993). Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. Proceedings of 31st Meeting of the ACL, Columbus, Ohio, USA, pp. 235-252.

Miyata, T., Utsuro, T. & Matsumoto, Y. (1997). Bayesian Network Models of Subcategorization and their MDL-Based Learning from Corpus. Proceedings of the 4th Natural Language Processing Pacific Rim Symposium, pp. 321-326.

Murphy, P.M. and Aha, D.W. (1993). UCI repository of machine learning databases. [Machine-readable data repository]. University of California, Department of Information and Computer Science, Irvine, CA.

Murphy, K.P. (1998). Inference and learning in hybrid Bayesian networks. Technical Report 990, UC Berkeley.

Narayanan, S. & Jurafsky, D. (1998). Bayesian Models of Human Sentence Processing. Proceedings of the 20th Annual Conference of the Cognitive Science Society, 752-757

Niyogi, S. (2002). Bayesian Learning at the Syntax-Semantics Interface. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia.

Partners of ESPRIT-291/860. (1986). Unification of the word classes of the ESPRIT Project 860. BU-WKL-0376, Internal Report.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann.

Schapire, R. (1999). Theoretical views of boosting and applications. In Proceedings of the Tenth International Conference on Algorithmic Learning Theory, pp. 13-25.

Skalak, D. (1994). Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms. Proceedings of the 11th Machine Learning Conference, 293-301, New Brunswick, Morgan Kaufmann.

Stamatatos, E., Fakotakis N. & Kokkinakis, G. (2000). A Practical Chunker for Unrestricted Text. Proceedings of the 2nd International Conference of Natural Language Processing, Patras, Greece, pp. 139-150.

Suzuki, J. (1993). A construction of Bayesian networks from databases on a MDL scheme. In Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, pp. 266-273.

Tasikas, A. (2002). Name Entity Recognition in Modern Greek Texts using Machine Learning, Diploma Thesis, University of Patras.

Tomek, I. (1976). Two modifications of CNN. IEEE Transactions on Systems, Man and Communications, SMC, 6, pp. 769-772.