Semi-Supervised Learning using Higher-Order Co-occurrence Paths to Overcome the Complexity of Data Representation

Murat Can Ganiz
Computer Engineering Department,
Faculty of Engineering
Marmara University, İstanbul, Turkey
murat.ganiz@marmara.edu.tr

***Abstract*— We present a novel approach to semi-supervised learning for text classification based on the higher-order co-occurrence paths of words. We name the proposed method as Semi-Supervised Semantic Higher-Order Smoothing (S3HOS). The S3HOS is built on a tri-partite graph based data representation of labeled and unlabeled documents that allows semantics in higher-order co-occurrence paths between terms (words) to be exploited. There are several graph-based techniques proposed in the literature to diffuse class labels from labeled documents to the unlabeled documents. In this study we propose a different and natural way of estimating class conditional probabilities for the terms in unlabeled documents without need to label the documents first. The proposed approach allows estimating class conditional probabilities for the terms in unlabeled documents and improve the estimation of terms in the labeled documents at the same time. We experimentally show that S3HOS can highly improve the parameter estimation and hence increase the classification accuracy particularly when the amount of the labeled data is scarce but unlabeled data is plentiful.**

Keywords— Semi-Supervised Learning, Naive Bayes, Semantic Smoothing, Higher-Order Naive Bayes, Text Classification.

I. Introduction

The semi-supervised classification algorithms aim to improve classifier performance in terms of accuracy by incorporating large amount of unlabeled data, which is readily available or easy to acquire. In many real-world applications it is costly to obtain labeled instances therefore the use of un-labeled instances to improve the accuracy may be a much cost effective alternative than labeling more instances. The research on semi-supervised learning (SSL) is of interest recent years and received considerable attention from the research community on several domains including image classification and text classification both of which are the domains that labeling instances are particularly expensive due to the size and complexity of their structures demands human intervention.

Several tutorials and surveys can be found about the semi-supervised learning in [15], [16], [17]. There are several different approaches to the semi-supervised learning, in particular semi-supervised classification. These include generative models, Semi-Supervised Support Vector Machines, graph-based models [24], co-training and multiview models [15]. In text classification domain, one of the most famous generative model is the semi-supervised variant of the Multinomial Naïve Bayes algorithm proposed by the Nigam, McCallum, and Mitchell [18]. This study uses a generative classifier; the Naïve Bayes (NB), to classify unlabeled instances starting with a small training set of labeled instances. This classification process of unlabeled instances is applied iteratively and optimized by using Expectation Maximization (EM). Our approach bears similarities with generative models since we use parameter estimations similar to the Naïve Bayes (NB). Additionally, it can be considered as a graph-based model as we too are using graph data structures. On the other hand, our approach diverges from generative models by not employing an iterative approach to label the unlabeled instances and it differs from the graph-based models as our graph data structures models terms (words) in documents instead of the documents. Furthermore, many graph-based SSL algorithms are intrinsically transductive [19] and as a result they can't naturally classify previously unseen instances that are not present during training.

In this paper, we present a novel approach to semi-supervised learning based on the higher-order co-occurrence paths of words. We name the proposed method as Semi-Supervised Semantic Higher-Order Smoothing (S3HOS). It is a higher-order learning approach. The higher-order learning approaches [3], [4], [13] exploits the implicit relations that stems from the basic structure of the natural language. If the terms or words are expressed together in a meaningful context such as a sentence or document, they obviously have a semantic relationship and these relationships are transitive which leads to the higher-order co-occurrences. These relations, specifically the higher-order co-occurrence relations link the terms in different contexts such as documents therefore blurring the documents borders. In our case, higher-order co-occurrence relations can link the terms in unlabeled documents to the ones in labeled documents.

More specifically, S3HOS is built on similar graph based data representation of the HOS [13], which allows semantics in higher-order co-occurrence paths between terms (words) to be exploited. Apart from the several graph-based techniques proposed in the literature to diffuse class labels from labeled documents to the unlabeled documents, we propose a novel and natural way of estimating class conditional probabilities for the terms in unlabeled documents without need to label them first. The approach allow estimating class conditional probabilities even for the terms in unlabeled documents and improve the estimation of terms in the labeled documents at the mean time.

We experimentally show that S3HOS can prominently improve the parameter estimation and increase the classification accuracy particularly when the amount of the labeled data is scarce and unlabeled data is plentiful.

The remainder of this paper is organized as follows: Section 2 presents background information about the studies of higher-order co-occurrence, Section 3 presents and analyzes the proposed classifier algorithm, Section 4 presents experimental setup and introduces datasets, Section 5 presents experiment results including some opinions and the last section presents a conclusion and future work.

II. BACKGROUND

Traditional machine learning algorithms assume that instances are independent and identically distributed (IID) [1]. There are several studies which exploit explicit link information in order to overcome the shortcomings of IID approach [1], [2] for supervised learning. However, the use of explicit links has a significant drawback; in order to classify a single instance, an additional context needs to be provided. In addition to these, there are many graph-based semi-supervised models, which relies on the explicit relationships generated by using a distance or similarity metric between documents [20], [21], [23], [24]. Similarly, these models are highly sensitive to the adjacency graph construction from documents and the regularization parameter parameters [22].

Latent Semantic Indexing (LSI) algorithm [6] is a widely used technique in Text Mining and Information Retrieval. It has been shown that LSI takes advantage of implicit higher-order (or latent) structure in the association of terms and documents. Higher-order relations in LSI capture “latent semantics” [7]. There are several LSI based classifiers. Among these, in [8] the authors propose a LSI based k-Nearest Neighborhood (LSI k-NN) algorithm in a semi-supervised setting for short text classification (SS-LSI-kNN).

In several previous studies, LSI motivated approaches are developed which explicitly make use of higher-order relations. These include a novel Bayesian framework called Higher-Order Naïve Bayes (HONB) [3], [4] and a novel data driven space transformation that allows vector space classifiers to take advantage of relational dependencies captured by higher-order paths between features [3] and Higher-Order Support Vector Machines (HOSVM) [5].

In general, the Naïve Bayes parameter estimation drastically suffers from sparsity due to the very large number of parameters to estimate in the text classification. The number of parameters corresponds to ($|V||C| + |C|$) where V denotes the dictionary and C denotes the set of class labels [10]. The sparsity of the text data increases the importance of smoothing methods since they are intended to distribute a certain amount of probability mass to the previously unseen events.

Based on the similar principles, Higher-Order Smoothing (HOS) for Naïve Bayes text classification takes the exploiting the higher-order relations concept one step further and exploit the relationships between instances of different classes in order to improve the parameter estimation [13]. The HOS is built on a graph-based data representation from the previous algorithms in higher-order learning framework such as HONB [4]. While, in the HONB, higher-order paths are extracted in the context of a particular class, in HOS, higher-order paths from the whole training set including all classes of documents. Consequently, it is possible to exploit the relations between the terms and the documents in the different classes. In other words, the basic units for parameter estimation, the higher-order paths, not only move beyond document boundaries but also the class boundaries to exploit the latent semantic information between terms. In this study, we develop the concept of higher-order paths even further to include unlabeled documents along with labeled dataset.

III. APPROACH

“Higher-order paths simultaneously capture term co-occurrences within documents as well as term sharing patterns across documents, and in doing so provide a much richer data representation than the traditional feature vector form” [3]. A higher-order co-occurrence path is shown in Fig.1. This figure depicts three documents, d_1, d_2 and d_3, each containing two terms t_1, t_2 in d_1, t_2, t_3 in d_2, and t_3, t_4 in d_3. Under the documents a small graph of four nodes and three edges is illustrated. This chain represents a higher-order path that links term t_1 with term t_4 through t_2 and t_3. This is an example of third-order path since three links, or “hops,” connect t_1 with term t_4. Similarly, there is a second-order path between t_1 and t_3 through t_2. The t_1 co-occurs with t_2 in document d_1, and t_2 co-occurs with t_3 in document d_2. Even if terms t_1 and t_3 never co-occur in any of the documents in a corpus, the regularity of these second order paths may reveal latent semantic relationships such as synonymy [13].

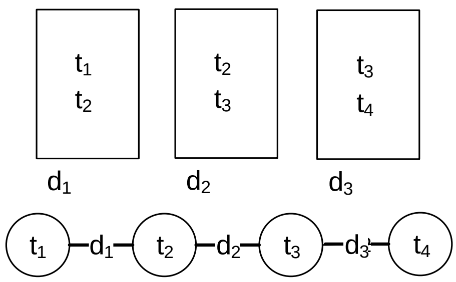

Fig. 1. Higher-Order co-occurrence. Adopted from [7]

In [13] a tri-partite graph is used to represent terms, documents, and class labels. In this study we use a similar tri-partite graph structure. However, in our case, the graph is created from the combination of labeled and unlabeled sets of documents. Therefore, as can be seen in Fig. 2, the set of terms and the set of documents consist of two regions demonstrating if they are from labeled or the unlabeled part of the dataset. The nodes representing documents consist of two separate node sets of labeled documents D_L and unlabeled documents D_U. Similarly, the nodes representing terms also divided into two sets T_U and T_L.

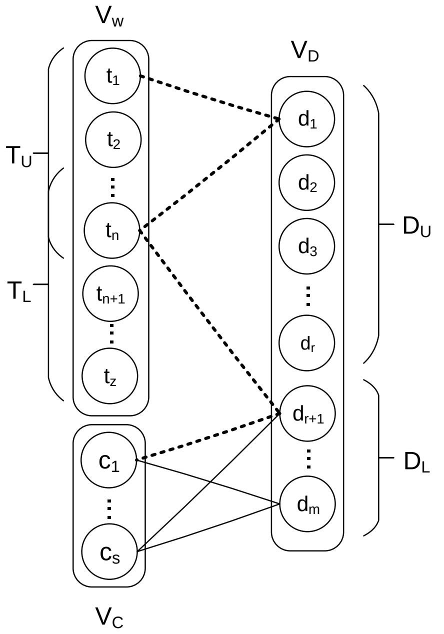

Fig. 2. Tripartite graph representation of labeled (D_L) and unlabeled (D_U) sets of documents, terms, and class labels, showing a second-order co-occurrence path between a term and a class label.

In general we use the tri-partite graph structure that is depicted in Fig. 2 and the following definitions in the following sections.

- D_U : the set of unlabeled documents
- D_L : the set of labeled documents
- T_U : the set of terms that occurs in unlabeled documents
- T_L : the set of terms that occurs in labeled documents
- T_{OU} : the set of terms that only occurs in unlabeled documents; $T_{OU} = T_U - T_L$
- T_{OL} : the set of terms that occurs only in labeled documents
- T_{UL}: the set of terms that occurs both in labeled and unlabeled documents; $T_{UL} = T_U \cap T_L$
- We assume that $D_U \cap D_L = \emptyset$ and $T_{UL} \neq \emptyset$.
- $(t_n,c_m)_1$ is the number of first-order paths between t_n and class label m (c_m). These are basically the co-occurrences, in binary form which correspond to the number of documents in class m that includes term t_n.
- $(t_n,c_m)_2$ indicates the number of second-order paths between t_n and class label m (c_m). These are higher-order co-occurrences.

t_1—d_1—t_n—d_{r+1}—c_1 is an example second-order co-occurrence path between t_1 and c_1 which is indicated by bold dashed lines in Fig.2. If we look carefully to this example, it is a special case which links a term that exist only in unlabeled documents to a class label since $t_1 \in T_{OU}$ and $t_n \in T_{UL}$ and $d_1 \in D_U$ and $d_{r+1} \in D_L$. By using this kind of relations we could be estimate parameters for the terms in unlabeled documents which is otherwise not possible without labeling them first.

In S3HOS, similar to the HOS [13], the parameters of the Naïve Bayes can be calculated from higher-order paths (i.e second-order paths) instead of documents as in Eq. 1 and 2.

$$P_1(t_n|c_m) = \frac{(t_n,c_m)_1}{\sum_i^{|T|}(t_i,c_m)_1} \quad (1)$$

$$P_2(t_n|c_m) = \frac{(t_n,c_m)_2}{\sum_i^{|T|}(t_i,c_m)_2} \quad (2)$$

In order to fuse the power of first-order and second-order paths, we can combine probability estimates from first-order paths and second-order paths using the linear interpolation as in Eq. 3. If the λ=1 we use only the second-order paths in parameter estimates. On the other hand if λ=0 we use only the first-order co-occurrences. It is important to note that λ also controls the effect of labeled documents in the overall estimate since only the terms in labeled documents have potential to have first-order co-occurrence with class labels. In other words, terms that occur only in unlabeled documents can't have a first-order co-occurrence with a class label.

$$P(t_n|c_m) = \frac{(1-\lambda)(t_n,c_m)_1 + \lambda(t_n,c_m)_2 + s}{(\sum_i^{|T|}(1-\lambda)(t_i,c_m)_1 + \lambda(t_i,c_m)_2) + 2s} \quad (3)$$

The s is the smoothing parameter in order to avoid zero probabilities. We assign a very small number ($0<s<<1$) in our experiments.

The second parameter of the Naïve Bayes; class prior probability for class m can be calculated using Eq. 4. We assume that each class includes at least one document.

$$P(c_m) = \frac{\sum_i^{|T|}(1-\lambda)(t_i,c_m)_1 + \lambda(t_i,c_m)_2}{\sum_j^{|C|}\sum_i^{|T|}(1-\lambda)(t_i,c_j)_1 + \lambda(t_i,c_j)_2} \quad (4)$$

This new formulation makes it possible to calculate class conditional probabilities for some of the terms that do not occur in documents of a particular class as well as the terms does not even occurs in the documents of the labeled dataset. Furthermore, the use of second-order paths provides more evidence for the terms that exist in a particular class by making use of the terms and documents in the unlabeled dataset.

The higher-order paths capture the term co-occurrences within a class of documents, term relation patterns across classes, and term relation patterns across labeled and unlabeled documents. Consequently, The Naive Bayes type probability estimation extracts patterns from a much richer data representation than the traditional vector space representation.

IV. Experiment Setup

We use three different datasets including widely used benchmark datasets in text classification domain. The first dataset is a widely used dataset in the text classification literature named WebKB[1]. The WebKB dataset includes web pages collected from computer science departments of different universities, organized in seven categories. We use four-class version of the WebKB dataset, which is used in previous works [13]. We call this dataset as WebKB4 in the following sections. WebKB4 dataset has a highly skewed class distribution. The second dataset, called 1150haber, includes 1150 news articles in five categories, (economy, magazine, health, politics and sports) collected from Turkish online newspapers [11]. Our third dataset; imdb is a movie reviews dataset that is commonly used in sentiment analysis studies [14] since each document is labeled as "negative" or "positive". The class label, document and vocabulary statistics are given in Table 1, where |C| is the number of classes, |D| is the number of documents, and |V| is the vocabulary size i.e. the number of distinct terms. All the datasets are preprocessed using stop word removal, stemming and reducing the dimensionality to 2000 terms using Information Gain (IG).

TABLE I. DESCRIPTIONS OF THE DATASETS BEFORE PREPROCESSING

Data Set	|C|	|D|	|V|
WebKB4	4	4,199	16,116
1150Haber	5	1150	11,038
imdb	2	2,000	16,679

We also provide term based statistics for each dataset in Table 2. As can be seen from this table, WebKB4 dataset diverges from others with much longer document sizes and the largest deviation of document lengths. These kind of divergences may have notable effect on the results of our model since the higher-order path based models fundamentally rely on the quantity and the quality of the shared terms between documents.

TABLE II. THE TERM BASED CHARACTERISTICS OF THE DATASETS

	Terms in documents		**Term length**		
Dataset	Avg	StdDev	Avg	StdDev	Sparsity
WebKB4	81.68	49.95	5.34	2.97	0.96
1150Haber	57.23	37.37	5.81	3.22	0.97
imdb	75.35	21.21	5.78	3.20	0.97

All these datasets are fully labeled i.e. each and every document has a single class label. In our experiments, in order to measure the performance of semi-supervised algorithms we divided a dataset into three chunks; training set, unlabeled set and test set by using several different percentages. Class labels of the unlabeled set are removed. The test set percentage is kept constant 20%. The supervised algorithms are trained with training set and evaluated using test set, while the semi-supervised algorithms are trained with both the training set and unlabeled set (includes no class labels), and evaluated with the test set, likewise. These percentages can be seen in Table 3.

TABLE III. LABELED AND UNLABELED DATA SPLIT PERCENTAGES

Training Set %	**Unlabeled Set %**	**Test Set %**
1	79	20
5	75	20
10	70	20
30	50	20
50	30	20
70	10	20

Ten folds are created for each dataset and each of six levels by randomly selecting training set percentage and the unlabeled set percentage of the data, and leaving the rest 20% for the test. The stratified sampling is used while dividing the datasets i.e. the class distributions is preserved. This approach is similar to [9] [12] [13]. Similar to the [13], the experiment results are presented using the average accuracy and standard deviations (StdDev) of these 10-fold result.

We employ up to second-order paths based on the experimental results of previous studies [3], [4], [7], [13].

We use several common semi-supervised text classifiers. These semi-supervised algorithms include semi-supervised Naïve Bayes (EM-NB) [18], and semi-supervised Latent Semantic Indexing k-Nearest Neighborhood (SS-LSI-kNN) [8] [4] [13]. We use the exact same parameter setting for the LSI-kNN ; The number of neighbors is set to 25 and *k* parameter of LSI is automatically optimized for each dataset and training set percentage by using the Kaiser approach. The SS-LSI-kNN algorithm is the semi-supervised variant of the LSI-kNN algorithm used in [13]. This is simply achieved by providing union of labeled dataset and unlabeled dataset as an input to the LSI algorithm. However the kNN part of the algorithm only considers the labeled instances in the LSI space which is generated using the union of labeled and large number of unlabeled instances. Due to the extremely computational and space high-complexity of the LSI algorithm and the extensive number of experiments we conduct (10 folds for each of the 6 training set percentages) for each dataset, we couldn't obtain SS-LSI-kNN results for our largest datasets of WebKB4. Our computational resources simply do not allow them to be finished on a reasonable time frame.

V. EXPERIMENT RESULTS AND DISCUSSION

In this section we provide a detailed empirical assessment of the proposed algorithm; Semi-supervised Semantic Higher-Order Smoothing (S3HOS) with several semi-supervised algorithms. Although we calculate quite a few different evaluation metrics such as F-measure (F1) and the area under the ROC curve (AUC), we use accuracy as our main evaluation metric in this section as it is the most common approach.

[1] http://www.cs.cmu.edu/~textlearning

In the following figures the algorithm we propose in this study is indicated with S3HOS accompanied by the λ (lambda) value which shows how the first-order and second-order paths are combined in parameter estimations. The values of λ are reported. λ = 1 means that only the second-order paths are used. On the other hand, λ = 0.5 means that first-order and second-order paths are combined with equal weights. Please see approach section for the details of the combination method. The figures compares S3HOS with the semi-supervised algorithms; semi-supervised Naïve Bayes (EM+MNB) and SS-LSI-kNN. We consider EM+MNB to be our baseline since it too is a Naïve Bayes based semi-supervised algorithm.

We start with the imdb dataset since we got the most interesting results from this dataset as can be seen in Fig.3. We can see that lambda parameter has a critical influence on the performance of S3HOS on this dataset. S3HOS with λ =1 achieves 88% accuracy that is more than 10% difference from its closest competitor when the training dataset size is only 1% and unlabeled data size is 79%. This is the highest gain we see among all our experimental results. Even the S3HOS, λ=0.5 performs better than our base-line EM+MNB through out the chart. Interestingly, the performance of S3HOS, λ=1 (second-order paths only) seems to decrease slightly as we increase the labeled dataset percentage. SS-LSI-kNN performs worse than other algorithms. This is usually the case throughout the experiments.

Our second dataset is 1150haber and in Fig. 4, the results are again quite interesting. The S3HOS λ=1 (second-order paths only) triumphs an accuracy of more than 90% only using the %1 of the data as labeled training set and using 79% of the data as unlabeled dataset. The accuracy reaches almost 95% when the training dataset size increased to 5%.

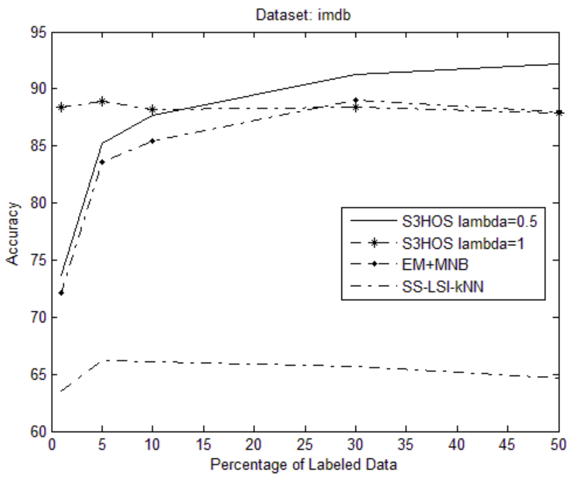

Fig. 3. Accuracy comparison of S3HOS with semi-supervised algorithms on imdb dataset.

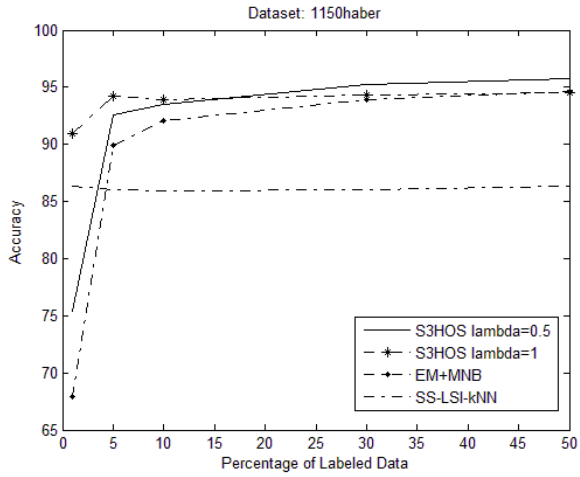

Fig. 4. Accuracy comparison of S3HOS with semi-supervised algorithms on 1150haber dataset.

S3HOS surpass our baseline semi-supervised algorithm EM+MNB on WebKB4 dataset as can be seen in Fig. 5. Interestingly, the lambda parameter seem to have a different effect in here. In other words, first-order paths seem to play a more important role in this dataset. We attribute this phenomenon to the highly skewed class distribution of the dataset. Furthermore, this dataset has the bigger documents compare to the others and the differences between the documents of different classes are relatively smaller. As a result, higher-order path may be introducing much more noise to the algorithm compare to the others.

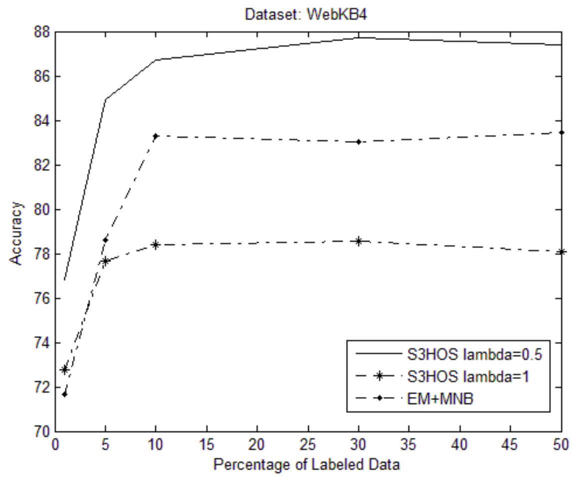

Fig. 5. Accuracy comparison of S3HOS with semi-supervised algorithms on WebKB4 dataset.

Overall, our experiments show that S3HOS outperforms our baseline semi-supervised model EM+MNB and SS-LSI-kNN (we could not provide results of the latter one for our larger dataset WebKB4 due to its exceptionally high complexity).

VI. CONCLUSIONS AND FUTURE WORK

We present a novel and a natural approach to semi-supervised learning for text classification based on the higher-order semantic smoothing (HOS) method that is developed in our previous study. We name the proposed method as Semi-Supervised Semantic Higher Order Smooth-ing (S3HOS). The S3HOS is built on a graph based data representation of the documents, terms and class labels, which allows semantics in higher-order co-occurrence paths between terms (and also the class labels) to be exploited. There are several graph-based techniques proposed in the literature to diffuse class labels from labeled documents to the unlabeled documents. In this study we propose a novel and natural way of estimating class conditional probabilities for the terms in unlabeled documents without need to label them first. The approach allow estimating class conditional probabilities even for the terms in unlabeled documents and improve the estimation of terms in the labeled documents at the mean time. This new formulation makes it possible to calculate class conditional probabilities for some of the terms that do not oc-cur in documents of a particular class as well as the terms does not even occurs in the documents of the labeled dataset. Furthermore, the use of second-order paths provides more evidence for the terms that exist in a particular class by making use of the terms and documents in the unlabeled dataset.

Higher-order paths capture the term co-occurrences within a class of documents, term relation patterns across classes, and term relation patterns across labeled and unlabeled documents. Consequently, The Naive Bayes type probability estimation extracts patterns from a much richer data representation than the traditional vector space representation.

We experimentally show that S3HOS can prominently improve the parameter estimation and increase the classification accuracy particularly when the amount of the labeled data is scarce and unlabeled data is plentiful.

ACKNOWLEDGMENT

This work is supported in part by The Scientific and Technological Research Council of Turkey (TÜBİTAK) grant number 111E239. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the TÜBİTAK.

REFERENCES

[1] Taskar B, Abbeel P, Koller D. Discriminative Probabilistic Models for Relational Data. In Proc. Uncertainty in Artificial Intelligence (UAI02), Alberta,Canada , August 1-4, 2002, pp. 485-492.

[2] Getoor L, Diehl C P. Link Mining: A Survey. ACM SIGKDD Explorations Newsletter, 2005, 7(2), pp. 3-12.

[3] Ganiz M C, Lytkin N, Pottenger W M. Leveraging Higher Order Dependencies between Features for Text Classification. In Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Bled, Slovenia, September 2009, pp. 375-390.

[4] Ganiz M C, George C, and Pottenger W M. Higher Order Naïve Bayes: A Novel Non-IID Approach to Text Classification. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7), pp. 1022-1034.

[5] Lytkin N. Variance-based clustering methods and higher order data transformations and their applications. Ph.D. Thesis, Rutgers University, NJ, U.S.A. (2009)

[6] Deerwester S C, Dumais S T, Landauer T K, Furnas G W, Harshman, R A. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 1990, pp. 391-407.

[7] Kontostathis A, Pottenger W M. A Framework for Understanding LSI Performance. Journal of the Information Processing and Management, 42(1), 2006, pp. 56-73.

[8] Sarah Z, Hirsh H. Transductive LSI for Short Text Classification Problems. In Proc. of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS), Miami Beach, FL, May 2004, PP. 556-561.

[9] McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. In Proc. AAAI 1998 Workshop on Learning for Text Categorization, Madison, Wisconsin, July 1998, pp. 41-48.

[10] McCallum A, Nigam K. Text classification by bootstrapping with keywords, EM and shrinkage. In Working Notes of ACL 1999 Workshop for the Unsupervised Learning in Natural Language Processing (ACL), Maryland, USA, pp. 52-58.

[11] Amasyalı M F, Beken A. Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması (Measurement of Turkish Word Semantic Similarity and Text Categorization Application). In Proc. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU) (Signal Processing and Communications Applications Conference), Kocaeli, Turkey, April 2009.

[12] Rennie J D, Shih L, Teevan J, Karger D R. Tackling the poor assumptions of naive bayes text classifiers. In Proc. ICML 2003, August, pp. 616-623.

[13] Poyraz, Mitat, Zeynep Hilal Kilimci, and Murat Can Ganiz. "Higher-Order Smoothing: A Novel Semantic Smoothing Method for Text Classification." Journal of Computer Science and Technology 29.3 (2014): 376-391.

[14] Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.

[15] Zhu, X. Semi-Supervised Learning Literature Survey, University of Wisconsin – Madison, Computer Sciences Technical Report, Last Modified on on December 14, 2007

[16] O Chapelle, B Schölkopf, A Zien, Semi-supervised learning. MIT Press, ISBN 978-0-262-03358-9, 2006.

[17] Zhu, X. (2010). Semi-supervised learning. In Encyclopedia of Machine Learning(pp. 892-897). Springer US.

[18] Nigam, K., McCallum, A., & Mitchell, T. (2006). Semi-supervised text classification using EM. Semi-Supervised Learning, 33-56.

[19] Krishnapuram, B., Williams, D., Xue, Y., Carin, L., Figueiredo, M., & Hartemink, A. J. (2004). On semi-supervised classification. In Advances in neural information processing systems (pp. 721-728).

[20] Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. The Journal of Machine Learning Research, 7, 2399-2434.

[21] Wang, F., & Zhang, C. (2008). Label propagation through linear neighborhoods. Knowledge and Data Engineering, IEEE Transactions on, 20(1), 55-67.

[22] de Sousa, C. A. R., Rezende, S. O., & Batista, G. E. (2013). Influence of graph construction on semi-supervised learning. In Machine Learning and Knowledge Discovery in Databases (pp. 160-175). Springer Berlin Heidelberg.

[23] Lebichot, B., Kivimäki, I., Françoisse, K., & Saerens, M. (2014). Semisupervised Classification Through the Bag-of-Paths Group Betweenness. Neural Networks and Learning Systems, IEEE Transactions on, 25(6), 1173-1186.

[24] Subramanya, A., & Talukdar, P. P. (2014). Graph-Based Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 8(4), 1-125.