*International Journal of Research in Advent Technology, Vol.3, No.12, December 2015*
*E-ISSN: 2321-9637*
*Available online at www.ijrat.org*

# Semi Supervised Text Classification with Universum

Miss.Gargi Joshi
*Department of Information Technology, D. Y. Patil College of Engineering Ambi, Pune*
*Email: joshigargi999@gmail.com*

**Abstract-** Semi-Supervised learning is a special form of classification. Traditional classifiers use only labelled data (feature / label pairs) to train. Labelled instances however are often difficult, expensive, or time consuming to obtain, as they require the efforts of experienced human annotators. Meanwhile unlabelled data may be relatively easy to collect, but there has been few ways to use them. We use unlabelled (random) data for categorization. Labelled data are hard to obtain while unlabelled data are abundant, therefore semi-supervised learning is a good idea to reduce human labour and improve accuracy. Semi-supervised learning addresses this problem by using large amount of unlabelled data, together with the labelled data, to build better classifiers. Because semi-supervised learning requires less human effort and gives higher accuracy, it is of great interest both in theory and in practice. Universum, a collection of non-examples that do not belong to any class of interest, has become a new research topic in machine learning. Text categorization does not exist then the huge data cannot be classified.The most of the existing systems in the market to do are Ada boost technique, naïve bayes, support vector machine, neural network, Particle Swaram Optimization (PSO),etc.. Fuzzy ANN and Bayesian probability algorithm are the process of semi-supervised text categorization and which results best compared to the existing techniques and are encouraging. Methodologies used in the system are Shannon info-gain , TF-IDF, K-means , Gaussian distribution , Fuzzy ANN , Bayesian probability , Atkinson index. Finally, the experiments use Reuters data set with several combinations. Experimental results indicate that the proposed algorithm can benefit from Universum examples and outperform several alternative methods, particularly when insufficient labeled examples are available. When the number of labeled examples is insufficient to estimate the parameters of classification functions, the Universum can be used to approximate the prior distribution of the classification functions. The experimental results can be explained using the concept of Universum introduced by Vapnik, that is, Universum examples implicitly specify a prior distribution on the set of classification functions. The key advantages of the proposed semi-supervised learning approach are: (a) performance improvement of any supervised learning algorithm with a multitude of unlabelled data, (b) efficient computation by the iterative boosting algorithm, and (c) exploiting both manifold and cluster assumption in training classification model.

**Index Terms-** Universum learning, text classification, machine learning, semi supervised learning

## 1. INTRODUCTION

Text Categorization has become important due to the strong growth in the volume of text documents available on the internet. Supervised learning is the main approach to this problem, and numerous state-of-the-art supervised learning algorithms have been proposed and successfully used in text categorization. However, problem in applying supervised learning methods to real-world problems is the cost of obtaining sufficient unlabeled training data, since supervised learning methods often require a large, prohibitive, number of unlabeled training examples to learn accurately. Labeling is time-consuming and typically done manually. Conversely, unlabeled data is relatively easy to collect, and many algorithms and experimental results have demonstrated that it can considerably improve learning accuracy in certain practical problems . Consequently, semi-supervised learning, which involves learning from a combination of both unlabeled and unlabeled data, has recently attracted significant interest .Semi-supervised learning often uses one of the following assumptions to model the structure of the underlying data distribution. First, the categorization function should be smooth with respect to the intrinsic structure revealed by known unlabeled and unlabeled data. Numerous studies view smoothness as an optimization with constraints problem, in which a regularization term is used for the problem at hand . Second, points on the same cluster or manifold frequently share the same label. Consequently, the categorization functions are naturally defined only on the sub-manifold in question rather than the total ambient space. Besides the above two assumptions, prior knowledge is another source of information that can improve categorization performance. While prior knowledge has proven useful for categorization it is notoriously difficult to apply in practice due to the difficulty of explicitly specifying prior knowledge. The Universum, introduced by Vapnik, provides a novel means to encode prior knowledge by giving examples. The Universum is a collection of examples that do not belong to any category of interest, but do belong to the

same domain as the problem. Weston devised an algorithm called U-supporting vector machine (SVM) to demonstrate that using Universum as a penalty term of the standard SVM objective function can enhance categorization performance., the Universum impacts semi-supervised learning with sufficient unlabeled examples. Both semi-supervised learning and learning with Universum rely on unlabeled examples to improve categorization performance. Learning with Universum and semi-supervised learning differ mainly in the distribution of unlabeled examples. While requiring unlabeled examples to share the same distribution as unlabeled ones, semi-supervised learning attempts to learn a model using a few unlabeled examples and numerous target unlabeled examples. Conversely, learning with Universum uses the examples with different distributions to the target ones, and aims to use Universum examples to estimate prior model information. Intuitively, the Universum examples should be close to the text categorization, since they do not belong to any class. Thus, the U-SVM is designed to minimize empirical categorization loss on the target examples rather than to give clear categorization assignments for the Universum examples. Most learning with Universum methods use margin to explain Universum and devise algorithms owing to the inspiration of U-SVM. Compared with most previous work on Universum, this investigation uses boosting technique to devise a semi-supervised learning with Universum algorithm. Analysis shows that using Universum can improve categorization performance, particularly when sufficient unlabeled examples are available. Although semi-supervised learning has achieved considerable success in the domain of machine learning, availability of only a few unlabeled examples may affect categorization performance. Universum can provide a means to model the prior of categorization functions, and the learning algorithm can obtain a text categorization by using Universum when sufficient unlabeled examples are available. The proposed system inspires us to design a semi-supervised learning algorithm with Universum to enhance semi-supervised learning categorization performance, particularly under conditions of sufficient unlabeled examples or data sets. To analyze the study findings, the experiments use different percentages of unlabeled examples to analyze the influence of Universum on categorization performance. Once more unlabeled examples become available, the benefits from Universum reduce. The experimental results relate the proposed method to Bayesian approaches. Finally, the experiments use four data sets with several combinations, and the experimental results indicate that the proposed algorithm can benefit from Universum examples. The proposed system consist of pre-given documents of unlabeled data sets , classify new documents , a standard categorization (unsupervised learning) problem.

## 2. RELATED WORK

Text categorization is an active research topic in the communities of Web data mining, information retrieval, and statistical machine learning. In the past decade, statistical learning techniques have been widely applied to text categorization [1], e.g, Bayesian classifiers [2], Support Vector Machines (SVM) [3], Logistic regression [4], and others. Empirical studies in recent years [5] have shown that SVM is the state-of-the-art technique. Traditional text categorization is conducted in the supervised setting, namely learning a classification model for text categorization from a pool of labeled documents. The supervised setting often requires a large amount of labeled documents before a reliable classification model can be built. Hence, an important research question in text categorization is how to build reliable text classifiers given a limited number of labeled documents. The key is to effectively explore unlabeled documents for text categorization. The first approach toward semi-supervised text categorization is multi-view learning. The main idea is to represent each document by multiple views and exploit unlabeled documents through the correlation among different views. This approach is especially effective for Web page and scientific document classification, in which the hyper-links between Web pages and the citation among research articles provide an additional representation for documents besides their textual contents [6,7,8]. Another example of multi-view learning is email categorization, in which the summaries of email texts [9] can be used as a complementary representation for emails. The co-training algorithm [10] and the EM algorithm for semi-supervised text categorization [11] also belong to this category. The second approach exploring unlabeled documents is to develop semi-supervised learning techniques that learn a classification model for text categorization from a mixture of labeled and unlabeled documents. The well-known examples within this category include Transductive SVM for text categorization [12,13]. The third approach is active learning [14,15,16] that aims to choose the most informative unlabeled documents for manually labeling. Finally, in addition to semi-supervised learning and active learning, another approach toward text categorization with small-size samples is to transfer the knowledge of a related text categorization task to the target text categorization task, which is closely related to transfer learning [17], domain adaptation [2], or transfer leaning from weakly-related unlabeled documents [11 ,3].

## 3. SEMI SUPERVISED LEARNING

In many machine learning applications, such as bioinformatics, web and text mining, text categorization, database marketing, spam detection, face recognition, and video indexing, abundant amounts of unlabelled data can be cheaply and automatically collected. However, manual labelling is often slow, expensive, and error-prone. When only a small number of labelled samples are available unlabelled samples could be used to prevent the performance degradation due to over fitting. The goal of semi-supervised learning is to employ a large collection of unlabelled data jointly with a few labelled examples for improving generalization performance.

Some semi-supervised learning methods are based on some assumptions that relate the probability $P(x)$ to the conditional distribution $P(Y = 1|X = x)$. Semi supervised learning is related to the problem of transductive learning. Two typical semi-supervised learning approaches are learning with the cluster assumption and learning with the manifold assumption. The cluster assumption requires that data within the same cluster are more likely to have the same label. The most prominent example is the transductive SVM

## 4. UNIVERSUM LEARNING

Universum data are given a set of unlabeled examples and do not belong to either class of the classification problem of interest. Contradiction happens when two functions in the same equivalence class have different signed outputs on a sample from the Universum. Universum learning is conceptually different from semi-supervised learning or transduction because the Universum data is not from the same distribution as the labeled training data. Universum learning implements a trade-off between explaining training samples (using large margin hyper planes) and maximizing the number of contradictions (on the Universum).

Universum is a new framework introduced by (Vapnik 1998) that is an alternative capacity concept to the large margin approach of SVMs. In this setting, one is given a set of labeled examples, and a collection of "non-examples" that do not belong to either class of interest. This collection, called the Universum, allows one to encode prior knowledge by representing meaningful concepts in the same domain as the problem at hand with examples.

Semi supervised and universum learning depend on unlabeled instances for classification. Semi supervised learning aims labeled and unlabeled instances universum has all distributions with historic knowledge and are hence close to decision boundary for instances which do not belong to a particular class of problem and work same as SVM .Universum works well when labeled instances are insufficient. Classification fails with insufficient labeled instances. Universum adjust the decision boundary by learning with insufficient A special kind of data is called universum [1], which does not belong to any classes of the problem at hand. [1] has shown that the universum data could boost the classification performance by encoding the prior knowledge of the domain. In addition, [4] and [15] studied the case that unlabeled data are a mixture of both relevant data, which are from the same domain as the current task, and irrelevant data, which are from a different task or the background. More specifically, [17] assumed the prior knowledge about the composition of the mixture, i.e., the universum data and the good quality same-domain data, is clear before learning a semi-supervised classification model instances

## 5. CONCLUSION

Universum set of data which does not belong to either of the two classes has become a leading research area in text classification and has a range of applications in the domain of bioinformatics, medical diagnosis, neural networks and text classification and can be combined with variety of approaches to improve classification performance by removing redundant data, reducing classification time, creating quality of clusters and derive best classification rules

## REFERENCES

[1] F. Sebastiani. Machine learning in automated text categorization. ACM Comput. Surv., 34(1):1–47, 2002.

[2 ]A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In Proceedings of the eleventh annual conference on Computational learning theory (COLT-1998), pages 92–100, New York, NY, USA, 1998. ACM Press.]

[3] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In ECML ’98: Proceedings of the 10th European Conference on Machine Learning, pages 137–142, London, UK, 1998. Springer-Verlag.

[4] J. Zhang, R. Jin, Y. Yang, and A. G. Hauptmann. Modified logistic regression: An approximation to svm and its applications in large-scale text categorization. In ICML ’03: Proceedings of the 20th international conference on Machine learning, pages 888–895, 2003.

[5] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. A comparison of implicit and explicit links for web page classification. In Proceedings of the 15th international conference on World Wide Web

(WWW 2006), pages 643–650, New York, NY, USA, 2006. ACM Press.

[6] Z. Xu, I. King, and M. R. Lyu. Web page classification with heterogeneous data fusion. In WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 1171–1172, New York, NY, USA, 2007. ACM Press.

[7] C. Li, J. R. Wen, and H. Li. Text classification using stochastic keyword generation. In ICML '03: Proceedings of the 20th international conference on Machine learning, pages 464–471, 2003.

[8] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. Machine Learning, 39(2-3):103–134, 2000.

[9] T. Joachims. Transductive inference for text classification using support vector machines. In ICML '99: Proceedings of the 16th international conference on Machine learning, pages 200–209, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[10] Z. Xu, R. Jin, J. Zhu, I. King, and M. R. Lyu. Efficient convex relaxation for transductive support vector machine. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, Advances in Neural Information Processing Systems 20, pages 1641–1648. MIT Press, Cambridge, MA, 2008

[11] D. J. C. MacKay. Information-based objective functions for active data selection. Neural Computation, 4(4):590– 604, 1992

[12] S. C. H. Hoi, R. Jin, and M. R. Lyu. Large-scale text categorization by batch mode active learning. In Proceedings of the 15th international conference on World Wide Web (WWW-2006), pages 633–642, New York, NY, USA, 2006. ACM Press.

[13] S. C. H. Hoi, M. R. Lyu, and E. Y. Chang. Learning the unified kernel machines for classification. In Proceedings of Twentith International Conference on Knowledge Discovery and Data Mining (KDD-2006), pages 187–196, New York, NY, USA, 2006. ACM Press.

[14] R. Caruana. Multitask learning. Machine Learning, 28(1):41–75

[15] H. Daum´e III. Frustratingly easy domain adaptation. In Conference of the Association for Computational Linguistics (ACL), Prague, Czech Republic, 2007.

[16] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Selftaught learning: transfer learning from unlabeled data. In ICML '07: Proceedings of the 24th international conference on Machine learning, pages 759–766, New York, NY, USA, 2007. ACM Press.

[17] E. Gabrilovich and S. Markovitch. Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. Journal of Machine Learning Research, 8:2297–2345, 2007.