Evaluating the Performance Estimators via Machine Learning Supervised Learning Algorithms for Dataset Threshold

Warda Imtiaz [a] Humaraia Abdul Ghafoor[b] Rabeea Sehar [c]

Tahira Mahboob[d] Memoona Khanum [e]
[d] Assistant Professor [e] Assistant Professor
[a, b, c] Department of Software Engineering Department of Computer Science
[a, b, c, d, e] Fatima Jinnah Women University, Pakistan.

ABSTRACT
Framework for user modeling is represented that is useful for both supervised and unsupervised machine learning techniques which will reduce the cost of development that is typically related to the knowledge-based approaches of machine learning for supervised approaches and user modeling that is basically required for the handling of the label-data. Experimental data is used for Research in bioinformatics. Vast amounts of experimental data populate the Current biological databases. Bioinformatics uses the machine learning concepts and has attained a lot of success in this research field. We focus on semi-surprised framework which incorporates labeled and unlabeled data in the general-purpose learner. Some of transfer graph, learning algorithms and the standard methods that include support vector machines and as a special case the regularized least squares can be obtained. We can use properties of reproducing the kernel Hilbert space to prove the new. Represented theorems provide the theoretical base for algorithms.

Keywords: Supervised and Unsupervised machine learning, intelligent analysis of data, techniques of data mining, evaluation of performance, bioinformatics, ensemble methods, User modeling, eye-tracking, graph transduction, and semi-supervised learning.

1. INTRODUCTION
For reducing the labeling effort for spoken language understanding used active and semi-supervised learning methods. Understanding the intent of the user can be framed as a classification problem in a goal-oriented call routing system.

Aim of Active learning is to minimize the number of labeled utterances which is by automatically selecting the utterance and these are most informative for labeling. One of the methods from these semi-supervised methods augments the training data for the unlabeled utterances by using machine labeled classes. Semi-Supervised learning approach is based on Gaussian random field model. In Weighted graph labeled and unlabeled data are represented as vertices, Similarity between instances encoding with edge weight. In terms of a Gaussian random field, the learning problem is then express on this graph. The mean of the field is characterized in terms of harmonic functions, and it is obtained by using matrix methods. The resulting learning algorithms have connections with random walks, spectral graph theory, and electric networks. In Machine Learning supervised learning is dominant methodology. In research, prediction is that supervised learning techniques are more powerful and suitable than the unsupervised techniques because for model optimization it provides the availability of labeled training data. In supervised learning risk minimization is presented as the suitable criteria to optimize. Basically supervised learning is used for implement risk minimization. Supervised learning techniques are widely used in multimedia data analysis.

2. SURVEY ON SUPERVISED LEARNING ALGORITHMS FOR DATASET THRESHOLD TO EVALUATE THE PERFORMANCE

2.1) Supervised Machine Learning: A Review of Classification Techniques (S. B. Kotsiantis)
The reason from on the surface supplied instances to produce general hypotheses [1] is done by supervised machine learning, which is the search for algorithms that make predictions about future instances. The distribution of class-label in term of predictor-features has goal of supervised learning that build a concise model.

When the values of the predictor features are known, but the value of the class label is unknown the resulting classifier is used to allocate class labels to the testing instances.

As a single article cannot give a complete review of all supervised machine learning classification algorithms this paper describes many supervised machine learning classification techniques .This technique used in guiding the researcher in interesting research directions and suggesting possible bias combinations that have yet to be explored, But we also hope that the references cited will cover the major theoretical issues.

2.2) Unsupervised and Supervised Machine Learning in User Modeling for Intelligent Learning Environments (S.Amershi and C.Conati)
Framework for user modeling is represented that is useful for both supervised and unsupervised machine learning techniques which will reduce the cost of development that is typically related to the knowledge-based approaches[2] of machine learning for supervised approaches and user modeling that is basically required for the handling of the label-data. Its result has shown that this approach is absolutely able to identify more complex data patterns that were initially found through the observation of data. Research paper established the new framework transferability across applications by comparing new results with the already existing results on the Applet.

2.3) An empirical comparison of supervised machine learning techniques in bioinformatics (A.C.Tan And D.Gilbert)
Experimental data is used for Research in bioinformatics. Vast

amounts of experimental data populate the Current biological databases. Bioinfomatics [3] uses the machine learning concepts and has attained a lot of success in this research field. As there are various algorithm are available researcher are facing the problem to choose the best algorithm. The empirical study is done on 4 various biological data sets, 9 different combined methods, 7 individual learning systems

and provide some suggested issues to answer the question. The questions are (1) how does one know which algorithm is suitable for their data set? (2) How does the effectiveness of a one algorithm is compared with the others? (3) Is single approach is better or combined. To classifying the biological data a comparison of different supervised machine learning techniques has been performed. Single methods could not consistently perform well over all the data sets. The nature of the training data decides the performance of the learning techniques. By experiments it is shown that combined methods are the better approach then single in terms of their specificity, positive predicted value, sensitivity and accuracy. Some rules-of-thumb has been suggested for the reader on choosing the best suitable machine learning algorithm for their dataset.

2.4) Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples(M.Belkin, P.Niyogi, and V.Sindhwani)

We proposed the family of learning algorithms that base on the new form of regularization which allow us to exploit geometry of marginal distribution[4]. We focus on semi-surprised framework which incorporates labeled and unlabeled data in the general-purpose learner. Some of transductive graph, learning algorithms and the standard methods that include support vector machines and as a special case the regularized least squares can be obtained. We can use properties of reproducing the kernel Hilbert space to prove the new. Represented theorems provide the theoretical base for algorithms. As a result (in the contrast of purely graph-base approach) we obtained the natural out-of-sample extensions to novel example and are able to handle the both transductive and the truly semi-surprise setting. We present the experimental evidence by suggesting that our semi-surprised algorithms be able to use the unlabeled data effectively. Finally have a brief description of the unsupervised and the fully supervised learning with in our general framework.

2.5) Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit (Shipeng Yu)

In this research paper, a probabilistic approach for supervised learning is described; when there have been multiple experts providing possibly noisy labels but no sheer gold standard. An estimate of the actual hidden labels [5] tells that the proposed algorithm evaluates the different experts. The proposed method is superior to the commonly used majority voting baseline indicated by experimental results. The area of computer-aided diagnosis (CAD) gives motivation for this research, where the task is to predict whether a suspicious region on a medical image is malignant or benign by building a classifier. A set of images is collected from different hospitals, in order to train such a classifier. Another scenario is the domain of text classification. To predict the category for a token of text is the task in this context. For ease of exposition the researchers used logistic regression. However, any generalized linear model or with any classifier that is trained with soft probabilistic labels; the proposed algorithm can be used.

2.6)Identifying Citing Sentences in Research Papers Using Supervised Learning (Kazunari Sugiyama Tarun Kumar Min-Yen Kan Ramesh C. Tripathi)

This paper is about using supervised learning approaches; with simple features extracted from research papers and developing a method for identifying citation sentences by constructing classifier. When research papers or articles are written, authors often make references to previous works in their own research field. Citations serve various purposes: as an acknowledgment of other's work, as evidence for claims, among other functions. The paper shows that approach, achieves a high level of accuracy [6], which constructs a supervised classifier from simple features. A system will take as input a research paper and identify the statements that require citation. Experimental results in this paper showed that effective features for training accurate models in both SVM and ME frameworks are; proper nouns and contextual classification of the previous and next sentence. According to an experiment, it is concluded that the accuracy of the task does not depend on classifiers; especially simple features bring better results among them. It may be difficult to construct an accurate classifier with very few overlapping features, as bigram features are often very sparse.

2.7) A Comparative Study of Training Algorithms for Supervised Machine Learning (Hetal Bhavsar, Amit Ganatra)

This research is a help to other researchers in studying the existing algorithms or developing innovative algorithms for applications requirements which are not available; related to the study of the existing classification algorithm and their comparison in terms of speed, accuracy, scalability and other issues. Data mining[7] provide large data set that involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships because; the tremendous amount of information stored in databases cannot only be used for further processing. Whereas these tools can include mathematical algorithm, statistical models, and machine learning method, Classification and prediction are two forms of data analysis that can be used to extract models: describing important data classes or to predict future data trends. Classification is also called supervised learning; it is capable of processing a wider variety of data than regression, as the instances are given with known labels. Labels are not known in unsupervised learning. Database instances and associated class label made the model from the training set this process is known as classification. Comparative study of some very well-known classification algorithms are concentrated in this research like; K-nearest neighbors, Decision Tree Induction, Neural Network, Bayesian Network, and Support Vector Machine.

2.8) Machine Learning Approach to Identifying the Dataset Threshold for the Performance Estimators in Supervised Learning (Zanifa Omary and Fredrick Mtenzi)

This paper presents the performance estimators in supervised machine learning experiments by a step by step guide for identifying the dataset threshold. In recent years, systems that can learn from experiences and adapt to their environments; are the goal of many researchers in different fields. Various algorithms such as decision trees, K-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Random Forests (RF)[8], has resulted into an establishment of this evolution t that they are transforming problems rising from industrial and scientific fields. Different performance measures and estimation methods, tend to perform differently when applied to different machine learning algorithms, based on the nature of the dataset; either balanced or unbalanced. The available performance measures are used while assessing and comparing one machine learning algorithm from the other such as precision, accuracy, error rate, f1-score, recall, and ROC analysis. There are various statistical tests, such as McNemar's test and a test of the difference of two proportions, also used to assess and compare classification algorithms, in addition to machine learning performance measures.

2.9) Supervised Learning (P ´adraig Cunningham, Matthieu Cord, and Sarah Jane Delany)

In Machine Learning supervised learning is dominant methodology. In research, prediction is that supervised learning techniques are more powerful and suitable than the

unsupervised[9] techniques because for model optimization it provides the availability of labeled training data. In supervised learning risk minimization is presented as the suitable criteria to optimize. Basically supervised learning is used for implement risk minimization. Supervised learning techniques are widely used in multimedia data analysis. The accessibility of annotated training data is the characteristic of supervised learning. The idea of 'supervisor' that show the learning system on the labels to relate with training. In classification problems these labels are basically class labels. These supervised learning techniques can be used for classify other unlabelled data. Analysis of supervised learning techniques is used on the theory of risk minimization. Two most important supervised learning techniques are used in multimedia research, also overview of nearest neighbor classifier and support vector machines.

2.10) Weakly Supervised Learning of Part-Based Spatial Models for Visual Object Recognition (David J. Crandall and Daniel P. Huttenlocher)

A new method of supervised learning for visual object recognition, information about class membership only provide from training data and object location or configuration are not provided, using these methods analysis of a model of local part appearance and a model of the spatial relations between those parts. A weakly supervised learning paradigm is not used for solving the problems of simultaneously learning appearance and spatial models. Some methods use a "bag" model in which only part appearance is considered and other methods used spatial models, only given the output of features. In previous techniques[10] appearance and spatial relations have instead used supervised learning process that provides extensive information about object part location. Weakly supervised technique provided better results than the previous highly supervised methods. Spatial models and richer appearance models are helpful in improving recognition performance. The results show that both spatial and appearance information can be helpful, the effect on performance depends largely on the particular object class and on the problem of the test dataset.

2.11) Semi-Supervised Learning Using Gaussian Fieldsand Harmonic Functions (XiaojinZhu,ZoubinGhahramani, JohnLafferty)

Semi-Supervised learning approach is based on Gaussian random field model. In Weighted graph labeled and unlabeled data are represented as vertices, Similarity between instances encoding with edge weight. In terms of a Gaussian random field, the learning problem is then express on this graph. The mean of the field is characterized in terms of harmonic functions, and it is obtained by using matrix methods. The resulting learning algorithms have connections with random walks, spectral graph theory, and electric networks. Methods to incorporate class priors and the predictions of classifier are discussed obtained by supervised learning. The experimental results[11] are presented for text and digit classification and representing that the structure of unlabeled data to improve classification accuracy. The framework is basically related to Gaussian process classification, and connection proposes principled ways of incorporating class priors.

2.12) Combining active and semi-supervised learning for spoken language understanding (Gokhan Tur , Dilek Hakkani-Tur , Robert E. Schapire)

For reducing the labeling effort for spoken language understanding used active and semi-supervised learning methods. Understanding the intent of the user can be framed as a classification problem in a goal-oriented call routing system. Aim of Active learning is to minimize the number of labeled utterances which is by automatically selecting the utterance and these are most informative for labeling. One of the methods from these semi-supervised methods augments the training data for the unlabeled utterances by using machine labeled classes. And the other second method augments the classification model[12] trained using human labeled utterance in a weighted manner with the machine labeled. Active and semi-supervised learning is combined using selectively sampled and automatically labeled data. Basically focus on which data will be label, and what to do with the remaining unlabeled data in active and semi-supervised learning methods. The active and semi–supervised learning methods presented can be easily applied to other statistical SLU and classification tasks.

Table 1: Machine Learning parameters, their meanings and possible values

Serial #	***Machine learning Parameters***	***Meanings***	***Possible values***
1	Error rate	Number of incorrect predictions against total predictions	Yes, No , and not defined
2	Precision	In information retrieval, where datasets are unbalanced	Yes, No , and not defined
3	Accuracy	Number of correct predictions over total number of predictions	Yes, No , and not defined
4	Recall	Proportion of number of items as positive to total	Yes, No , and not defined
5	ROC (Receiver operating characteristics) factor	Graph visualization, selecting, and organizing classifier based on their performance	Yes, No , and not defined
6	Case study	Reference to some experimentation	Yes, No , and not defined

Table 2: Quality parameters, their meanings and possible values

Serial #	***Quality Parameters***	***Meanings***	***Possible Values***
1	Extendibility	New features can be easily adapted by a system or techniques	Yes, No , and not defined
2	Performance	Recovery rate of the system	Yes, No , and not defined
3	Integrity	Protection against unauthorized access protecting software and programs	Yes, No , and not defined
4	Reusability	How well the modules of techniques can be reused in a new system	Yes, No , and not defined
5	Robustness	Appropriate performance of a technique under cases not covered by the requirement. This is complementary to correctness.	Yes, No , and not defined
6	Efficiency	While using minimum resources to produces high throughput	Yes, No , and not defined
7	Complexity	To what extent modules are inter-related	Yes, No , and not defined
8	Security	no one can access the personal information of a specific person	Yes, No , and not defined
9	Cost effective	Relation of cost to the productive process	Yes, No , and not defined
10	Reliability	To perform functions without experiencing failure.	Yes, No , and not defined

Table 3: Analysis Table of parameters against authors

#Serial	Authors	RateError	Precision	Accuracy	Recall	FactorROC	Extendibility	Performance	Integrity	Reusability	Robustness	Efficiency	Complexity	Effective Cost	Reliability	StudyCase
1	S. B. Kotsiantis	Y	Y	Y	ND	Y	Y	Y	ND	Y	N	ND	Y	N	Y	Y
2	S.Amershi and C.Conati	N	Y	ND	Y	Y	ND	ND	Y	Y	Y	Y	N	Y	ND	N
3	A.C.Tan And D.Gilbert	ND	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	N	ND	Y	Y
4	M.Belkin, P.Niyogi, and V.Sindhwani	Y	Y	N	Y	Y	Y	Y	ND	Y	ND	ND	Y	Y	Y	N
5	V. C. Raykar, S. Yu, L. H. Zhao	N	Y	ND	N	Y	Y	ND	N	N	Y	Y	N	Y	Y	Y
6	K. Sugiyama, T. Kumar, M. Kan	Y	N	Y	N	Y	ND	N	Y	Y	N	ND	Y	N	Y	N
7	H. Bhavsar and A. Ganatra	Y	Y	N	Y	N	Y	Y	N	Y	Y	ND	N	N	ND	Y
8	Z. Omary and F. Mtenzi	ND	Y	Y	ND	Y	N	N	Y	ND	Y	N	Y	Y	N	Y
9	P.Cunningham, M.Cord, and S. J.Delany	Y	Y	Y	Y	ND	Y	N	N	Y	ND	Y	N	Y	N	Y
10	G. Tur, D.Tur, and R.E. Schapire	N	N	Y	N	Y	ND	Y	Y	Y	N	Y	N	Y	N	N
11	D.J. Crandall, D. P. Huttenlocher	ND	Y	Y	ND	N	Y	Y	ND	Y	Y	Y	Y	Y	Y	Y
12	X.Zhu,Z.Ghahramani,a ndJ.Lafferty	Y	Y	Y	Y	ND	Y	N	N	Y	Y	ND	Y	N	N	Y

Yes=Y, No=, Not Defined=ND

3. CONCLUSION

The comparison of the most well-known classification algorithms like neural network, and Bayesian network, decision trees, nearest neighbor and support vector machine has been analyzed and compared in detail. The aim behind this study was to search the key ideas of these algorithms and in result find out the current research issues. This work can help other researchers who are doing an advanced course on classification in machine learning. The study of comparative analysis of theses algorithm had shown that each has its own set of advantages and disadvantages, as well as its own area of implementation. A single algorithm cannot satisfy all the criteria of classifications. One could create a classifier which can be developed by the integration of two or more algorithms; by combining their strength they can best satisfy the criteria.

4. REFERENCES

[1] S. B. Kotsiantis. "2007 Supervised Machine Learning: A Review of Classification Techniques" in Informatica 31 (2007) 249-268 249. Avaiable: http://www.informatica.si/PDF/31-3/11_Kotsiantis%20-%20Supervised%20Machine%20Learning %20-%20A%20Review%20of...pdf

[2] S.Amershi and C.Conati. "2007 Unsupervised and Supervised Machine Learning in User Modeling for Intelligent Learning Environments " in IUI'07, January 28–31, 2007, Honolulu, Hawaii, USA. Available: https://www.cs.ubc.ca/~conati/my.../IUI07-10604SaleemaCAMERA.pdf

[3] A.C.Tan And D.Gilbert. "2003 An empirical comparison of supervised machine learning techniques in bioinformatics" in the Proceedings of the First Asia Pacific Bioinformatics Conference (APBC 2003). Available: core.ac.uk/download/pdf/335643.pdf

[4] M.Belkin, P.Niyogi, and V.Sindhwani. "2006 Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples." In Journal of Machine Learning Research 7 (2006) 2399-2434. Available: vikas.sindhwani.org/MR.pdf

[5] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebk, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. "Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit." 26th International Conference on Machine Learning, Montreal, Canada, 2009. Available: http://facweb.cti.depaul.edu/research/vc/seminar/Paper/37.pdf

[6] K. Sugiyama, T. Kumar, M. Kan, and R. C. Tripathi. "Identifying Citing Sentences in Research Papers Using Supervised Learning." Media Development Authority (MDA) grants "Interactive Media Search," R-252-000-325-279. Available: http://www.ijsce.org/attachments/File/v2i4/D0887072412.pdf

[7] H. Bhavsar and A. Ganatra. "A Comparative Study of Training Algorithms for Supervised Machine Learning." International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4, September 2012. Available: http://www.ijsce.org/attachments/File/v2i4/D0887072412.pdf

[8] Z. Omary and F. Mtenzi. "Machine Learning Approach to Identifying the Dataset Threshold for the Performance Estimators in Supervised Learning." International Journal for Infonomics (IJI), Volume 3, Issue 3, September 2010.

[9] P.Cunningham, M.Cord, and S. J.Delany.(2008). "Supervised Learning".Springer. [On- line].16, pp. 289. Available: http://www.springer.com/978-3- 540-75170-0[2008].

[10] G. Tur, D.Tur, and R.E. Schapire.(2005). "Combining active and semi-supervised learning for spoken language understanding".Elsevier. [On- line]. 45. (2005),pp. 171–186. . Available:www.Sciencedirect.com.

[11] D.J. Crandall, D. P. Huttenlocher. "Weakly Supervised Learning of Part-Based Spatial Models for Visual Object Recognition." Internet: www.cs.cornell.edu/~dph/papers/eccv06- unsup.pdf.

[12] X.Zhu,Z.Ghahramani,andJ.Lafferty."Semi Supervised Learning Using Gaussian Fields and Harmonic Functios." Internet: mlg.eng.cam.ac.uk/zoubin/papers/zgl.pdf.