DATA SCIENCE INCEPTION OF AN IDEA TO PUBLICATION

Dr. Saptarsi Goswami সপ্তর্ষি গোস্বামী

Assistant professor – comp sc. Bangabasi Morning College

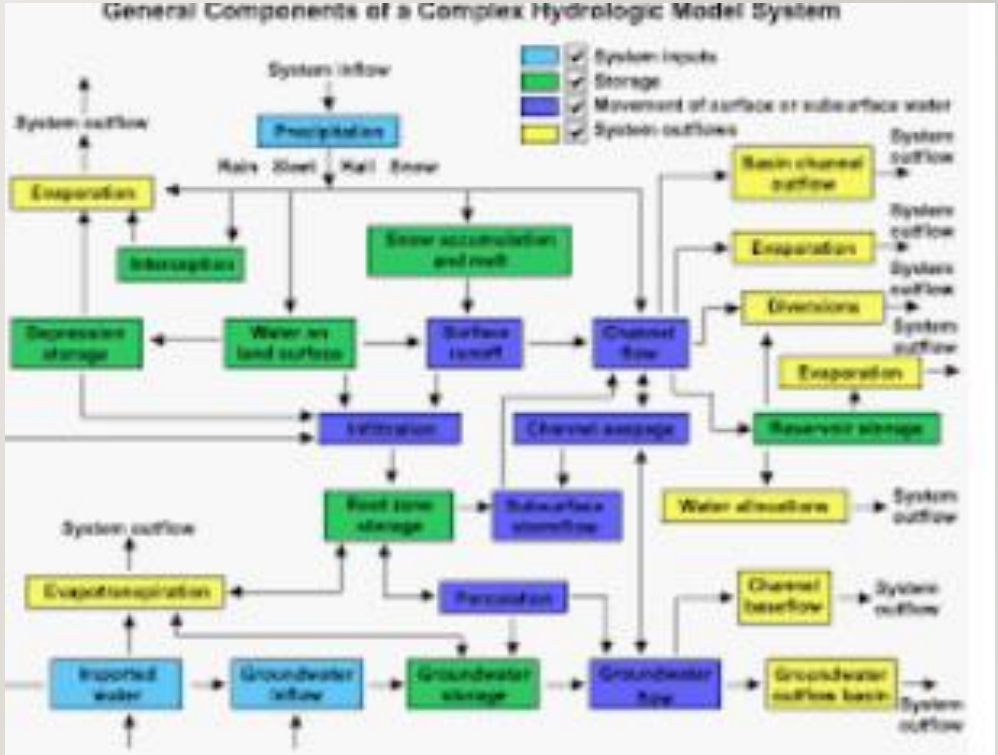
S4DS executive committee member

ODSC kolkata chapter LEAD

CONTENT

- Types of Research
- Types of Articles
- How to choose one

STARTING POINT



SYSTEMATIC LITERATURE STUDY

- Where to start
  - IEEE, ACM, Springer, Elsevier
  - Google Scholar
  - ResearchGate

TYPES

- Publishing at : Conference, Journal
- Length : Short (Conference Papers, Posters, Letters), Full Length
- Indexing : Scopus, DBLP, Web of Science
- Type of Research
  - ✓ Review Paper
  - ✓ Empirical Study
  - ✓ Contributions

TYPES OF PAPER

- Review Paper (Excerpt from our paper)

Abstract:
Healthcare is a fast-growing field in developed and developing countries alike. Indian healthcare has also witnessed rapid growth in the recent past. The data generated is of very high volume and exhibit wide diversity, and hence data mining opportunities available in various areas of health care industry are immense. In this paper, a brief review of data mining applications in healthcare has been presented. The key differentiating factors of the paper is its focus on India and a patient lifecycle oriented view of the problems. As found by the study, most of the researches have been focused on predicting a disease which is a part of ongoing care management. This paper presents many uncharted areas of future research in this domain especially in Indian context.

Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM computing surveys (CSUR). 2009 Jul 30;41(3):1-58. - 7000 Citations

ONE FUNDAMENTAL QUESTION HOW THIS IS DIFFERENT

The existing surveys discuss anomaly detection techniques that detect the simplest form of anomalies. We distinguish simple anomalies from complex anomalies. The discussion of applications of anomaly detection reveals that for most application domains, the interesting anomalies are complex in nature, while most of the algorithmic research has focussed on simple anomalies.

Table I. Comparison of our Survey to Other Related Survey Articles. 1—Our Survey, 2—Hodge and Austin [2004], 3—Agyemang et al. [2006], 4—Markou and Singh [2003a], 5—Markou and Singh [2003b], 6—Patcha and Park [2007], 7—Beckman and Cook [1983], 8—Bakar et al. [2006]

		1	2	3	4	5	6	7	8
Techniques	Classification Based	√	√	√	√		√		
	Clustering Based	√	√	√			√		
	Nearest Neighbor Based	√	√	√			√		√
	Statistical	√	√	√		√	√	√	√
	Information Theoretic	√							
	Spectral	√							
Applications	Cyber-Intrusion Detection	√					√		
	Fraud Detection	√							
	Medical Anomaly Detection	√							
	Industrial Damage Detection	√							
	Image Processing	√							
	Textual Anomaly Detection	√							
	Sensor Networks	√							

SYSTEMATIC CLASSIFICATION

Table 2: Various data types as used in the papers

Data Type	References
Structured	[14-19], [21-23], [25-31], [33-45] , [52]
Unstructured	
Text	[49],[50]
Image	[32] ,[34],[50],[53]
Time Series	[24],[54-57]

Table 4: Popular algorithms

Data Mining Task	References
Apriori	[38],[52]
Artificial Neural Network	[14],[17],[19],[21],[27],[28], [32],[33],[34],[36],[37],[39], [43],[45],[47],[48]
Decision Trees (Includes J48, CART, C 4.5)	[15],[16],[21],[22],[27],[28], [31],[34],[35],[43],[45],[49], [50],[52],[53],[56]
Association Rule Mining	[25],[38],[52]
K-Means	[18],[55]
kNN	[22],[25],[31]
Linear Regression	[44],[50]
Logistics Regression	[36],[39],[45]
Naïve Bayes	[15],[22],[23],[26],[27],[29],[31],[37]
SVM	[21],[27],[29-31],[36],[37],[41], [42],[56]

NEW DOMAIN (8800 CITATION)

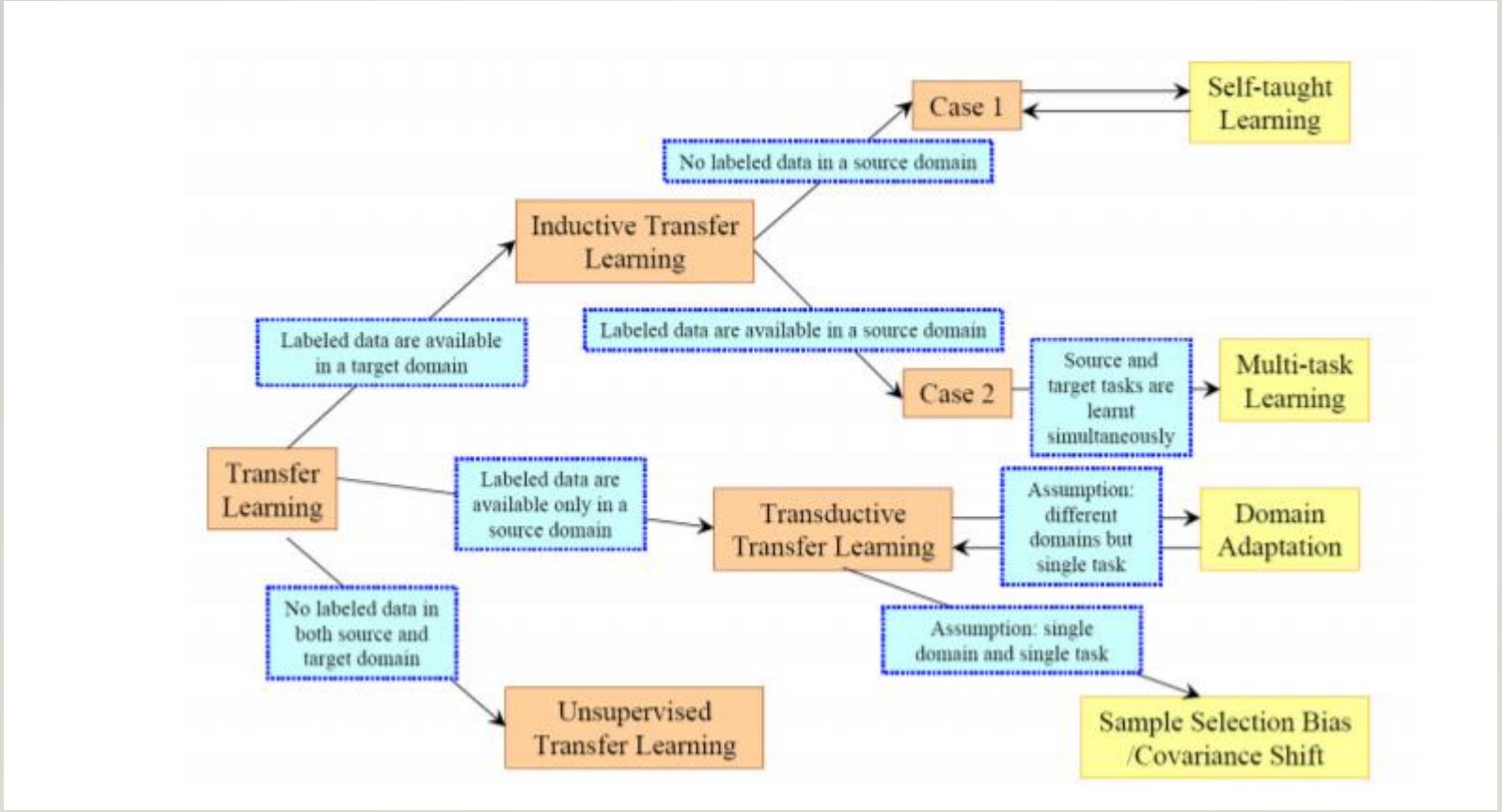
A Survey on Transfer Learning

Sinno Jialin Pan and Qiang Yang *Fellow, IEEE*

Abstract—A major assumption in many machine learning and data mining algorithms is that the training and future data must be in the same feature space and have the same distribution. However, in many real-world applications, this assumption may not hold. For example, we sometimes have a classification task in one domain of interest, but we only have sufficient training data in another domain of interest, where the latter data may be in a different feature space or follow a different data distribution. In such cases, knowledge transfer, if done successfully, would greatly improve the performance of learning by avoiding much expensive data labeling efforts. In recent years, transfer learning has emerged as a new learning framework to address this problem. This survey focuses on categorizing and reviewing the current progress on transfer learning for classification, regression and clustering problems. In this survey, we discuss the relationship between transfer learning and other related machine learning techniques such as domain adaptation, multi-task learning and sample selection bias, as well as co-variate shift. We also explore some potential future issues in transfer learning research.

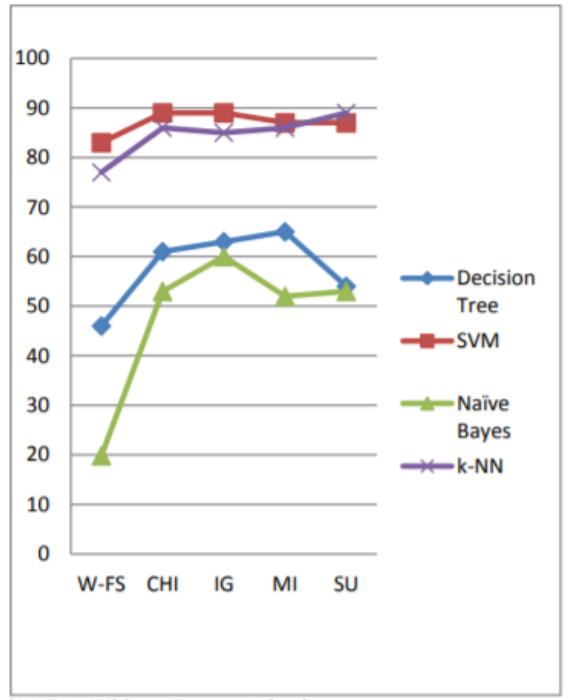
Index Terms—Transfer Learning, Survey, Machine Learning, Data Mining.

TAXONOMY


Transfer Learning
Labeled data are available in a target domain
Inductive Transfer Learning
No labeled data in a source domain
Case 1
Self-taught Learning
Labeled data are available in a source domain
Case 2
Source and target tasks are learnt simultaneously
Multi-task Learning
Labeled data are available only in a source domain
Transductive Transfer Learning
Assumption: different domains but single task
Domain Adaptation
Assumption: single domain and single task
Sample Selection Bias /Covariance Shift
No labeled data in both source and target domain
Unsupervised Transfer Learning

SECOND TYPE (EMPIRICAL)

occasions. Another finding of the study is, naïve bayes was the worst in terms of accuracy with an average prediction of rate of 24.2% and with feature selection there is a dramatic improvement and the result is then comparable to other classifiers with reduced set of features. There will be a need to work further on theoretical foundations to make it more appropriate for text classification.



*W-FS= Without Feature Selection

Figure-1 Performance of set accuracy rate involved in CNAE-9 dataset

EXAMPLE 1

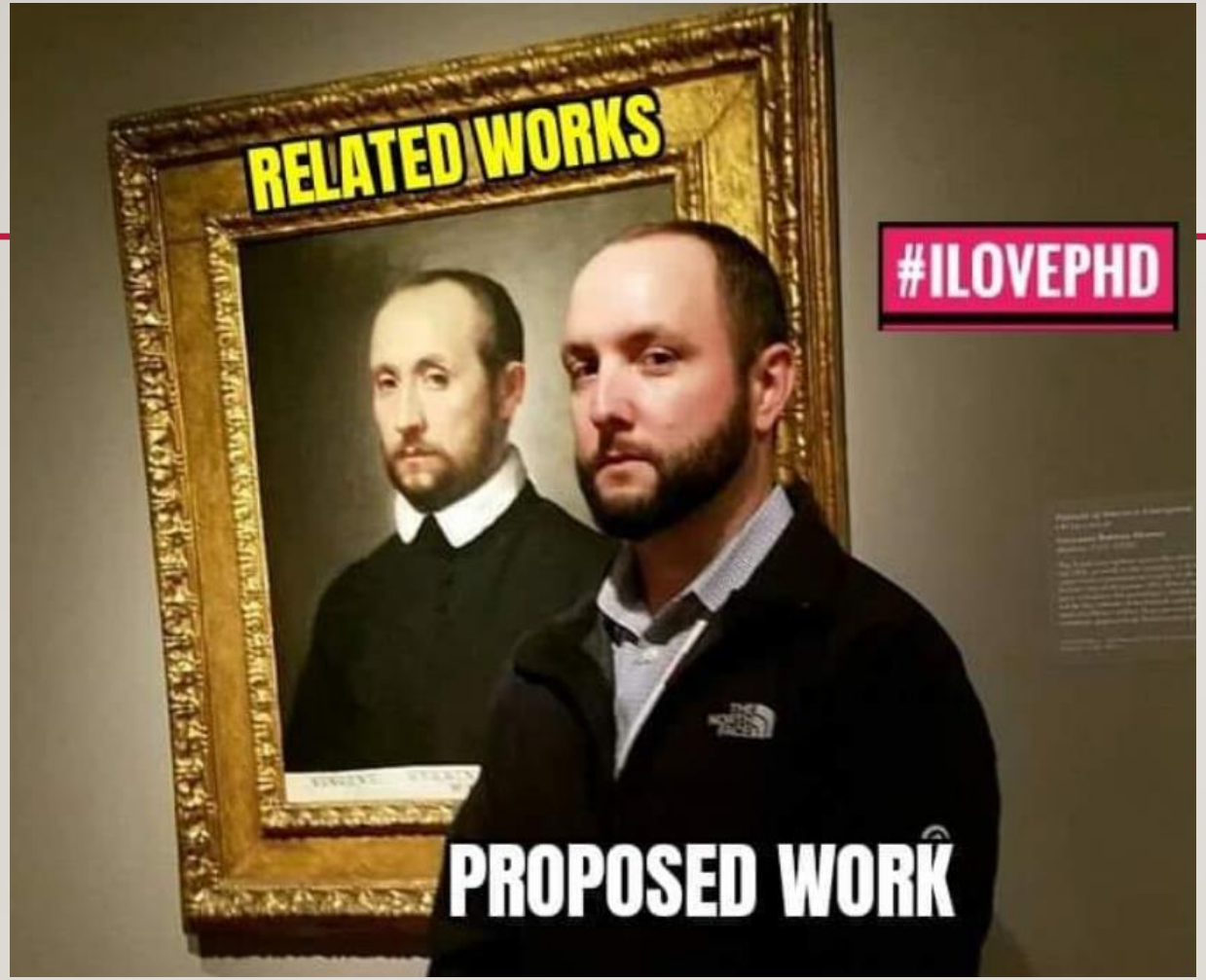
Abstract

Opinion mining deals with determining of the sentiment orientation—positive, negative, or neutral—of a (short) text. Recently, it has attracted great interest both in academia and in industry due to its useful potential applications. One of the most promising applications is analysis of opinions in social networks. In this paper, we examine how classifiers work while doing opinion mining over Spanish Twitter data. We explore how different settings (n-gram size, corpus size, number of sentiment classes, balanced vs. unbalanced corpus, various domains) affect precision of the machine learning algorithms. We experimented with Naïve Bayes, Decision Tree, and Support Vector Machines. We describe also language specific preprocessing—in our case, for Spanish language—of tweets. The paper presents best settings of parameters for practical applications of opinion mining in Spanish Twitter. We also present a novel resource for analysis of emotions in texts: a dictionary marked with probabilities to express one of the six basic emotions(Probability Factor of Affective use (PFA)(Spanish Emotion Lexicon that contains 2,036 words.

EXAMPLE 2

Abstract:

This paper discusses a comprehensive suite of experiments that analyze the performance of the random forest (RF) learner implemented in Weka. RF is a relatively new learner, and to the best of our knowledge, only preliminary experimentation on the construction of random forest classifiers in the context of imbalanced data has been reported in previous work. Therefore, the contribution of this study is to provide an extensive empirical evaluation of RF learners built from imbalanced data. What should be the recommended default number of trees in the ensemble? What should the recommended value be for the number of attributes? How does the RF learner perform on imbalanced data when compared with other commonly-used learners? We address these and other related issues in this work.


RELATED WORKS
#ILOVEPHD
PROPOSED WORK

CONTRIBUTION PAPER (DAIVID M BLEI) LDA 33000

In this paper we consider the problem of modeling text corpora and other collections of discrete data. The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.

- In the popular tf-idf scheme (Salton and McGill, 1983), a basic vocabulary of "words" or "terms" is chosen, and, for each document in the corpus, a count is formed of the number of occurrences of each word. the approach also provides a relatively small amount of reduction in description length and reveals little in the way of inter- or intra document statistical structure
- LSI uses a singular value decomposition of the X matrix to identify a linear subspace in the space of tf-idf features that captures most of the variance in the collection. This approach can achieve significant compression in large collections. can capture some aspects of basic linguistic notions such as synonymy and polysemy
- In pLSI, each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers.

ADDRESSING THE ITEM COLD-START PROBLEM BY ATTRIBUTE-DRIVEN ACTIVE LEARNING

Abstract:

In recommender systems, cold-start issues are situations where no previous events, e.g., ratings, are known for certain users or items. In this paper, we focus on the item cold-start problem. Both content information (e.g., item attributes) and initial user ratings are valuable for seizing users' preferences on a new item. However, previous methods for the item cold-start problem either (1) incorporate content information into collaborative filtering to perform hybrid recommendation, or (2) actively select users to rate the new item without considering content information and then do collaborative filtering. In this paper, we propose a novel recommendation scheme for the item cold-start problem by leveraging both active learning and items' attribute information. Specifically, we design useful user selection criteria based on items' attributes and users' rating history, and combine the criteria in an optimization framework for selecting users. By exploiting the feedback ratings, users' previous ratings and items' attributes, we then generate accurate rating predictions for the other unselected users. Experimental results on two real-world datasets show the superiority of our proposed method over traditional methods.

Towards better exploiting convolutional neural networks for remote sensing scene classification

Abstract

We present an analysis of three possible strategies for exploiting the power of existing convolutional neural networks (ConvNets or CNNs) in different scenarios from the ones they were trained: full training, fine tuning, and using ConvNets as feature extractors. In many applications, especially including remote sensing, it is not feasible to fully design and train a new ConvNet, as this usually requires a considerable amount of labeled data and demands high computational costs. Therefore, it is important to understand how to better use existing ConvNets. We perform experiments with six popular ConvNets using three remote sensing datasets. We also compare ConvNets in each strategy with existing descriptors and with state-of-the-art baselines. Results point that fine tuning tends to be the best performing strategy. In fact, using the features from the fine-tuned ConvNet with linear SVM obtains the best results. We also achieved state-of-the-art results for the three datasets used.

TYPICAL SECTIONS

- Abstract
- Introduction: Overview of the field, Importance, Establishing motivation with a high level summary
- Literature Review :More thorough and structured review
- Proposed Method:
- Materials and Method: Data Collection, Hardware, Software, parameters, assumptions
- Results and Discussion:
- Conclusion
- References


HIGH
IMPACT
PhD
MEMES
I cite a guy who cited a guy...
Who cited another guy...

CHOOSING JOURNALS, CONFERENCES

- Editorial Board / Technical Committee
- Volume / Edition
- Indexing

ধন্যবাদ