LEARNING PROXIMITY RELATIONS

FOR FEATURE SELECTION

**ABSTRACT**

This work presents a feature selection method based on proximity relations learning. Each single feature is treated as a binary classifier that predicts for any three objects X, A, and B whether X is close to A or B. The performance of the classifier is a direct measure of feature quality. Any linear combination of feature-based binary classifiers naturally corresponds to feature selection. Thus, the feature selection problem is transformed into an ensemble learning problem of combining many weak classifiers into an optimized strong classifier. We provide a theoretical analysis of the generalization error of our proposed method which validates the effectiveness of our proposed method. Various experiments are conducted on synthetic data, four UCI data sets and 12 microarray data sets, and demonstrate the success of our approach applying to feature selection. A weakness of our algorithm is high timecomplexity.

**INTRODUCTION:**

IN many pattern recognition and machine learning applications, such as appearance-based image classification, document clustering, data mining, and information retrieval, we are involved to deal with high-dimensionality data which has thousands of features. Learning and classifying in such a high-dimensionality space is extremely difficult due to the curse of dimensionality.In fact, a small fraction among thousands of features is significant and relevant to their classes. The remaining is insignificant which only complicates data learning and modeling. Those insignificant features may seriously degrade the performance of machine learning algorithms. When involved with many insignificant features, even Support Vector Machine (SVM),as one of the most successful classifiers, alsoworks badly in that situation .Thus, those insignificant features are in a way, irrelevant, redundant, and need to be removed.the physical meaning loss, feature selection aims to select a feature subset which can highly preserve the original properties of samples for learning tasks. Beyond alleviating the effect of curse of dimensionality and speeding up the learning processing, there is a distinguishing advantage for feature selection. It’s beneficial to discover the potential association among samples by visualizing the original attributes of points in low-dimension. A typical application for feature selection is Microarray Analysis. Many researchers have explored the microarray technology to build cancer diagnosis, prognosis and prediction from gene expression data. However, the number of gene from microarray data is significantly large, and each gene carries independent genetic instructions for the development of the living organisms. Discovering the underlying associations from gene expressions to cancers needs feature selection techniques not only to reduce efficiently the high-dimensionality gene expression data, but also to preserve the physical integrity of gene for subsequent biological analysis.

**LITERATURE SURVEY:**

Title : Dynamic Infinite Relational Model for Time-varying Relational Data

Analysis.

Author : Katsuhiko Ishiguro, Tomoharu Iwata Naonori Ueda, Joshua Tenenbaum.

Year : 2008

**Description:**

We propose a new probabilistic model for analyzing dynamic evolutions of relational data, such as additions, deletions and split & merge, of relation clusters like communities in social networks. Our proposed model abstracts observed time- varying object-object relationships into relationships between object clusters. We extend the infinite Hidden Markov model to follow dynamic and time-sensitive changes in the structure of the relational data and to estimate a number of clusters simultaneously. We show the usefulness of the model through experiments with synthetic and real-world data sets.

**LITERATURE SURVEY:**

Title : A probabilistic model of cross-categorization

Author : Patrick Shafto a, Charles Kemp b, VikashMansinghka , Joshua

Year : 2012

**Description:**

Most natural domains can be represented in multiple ways: we can categorize foods in terms of their nutritional content or social role, animals in terms of their taxonomic groupings or their ecological niches, and musical instruments in terms of their taxonomic categories or social uses. Previous approaches to modeling human categorization have largely ignored the problem of cross-categorization, focusing on learning just a single system of categories that explains all of the features. We also formalize two commonly proposed alternative explanations for cross-categorization behavior: a features-first and an objects-first approach. The features- first approach suggests that cross-categorization is a consequence of attentional processes, where features are selected by an attentional mechanism first and categories are derived second. The objects-first approach suggests that cross-categorization is a consequence of repeated, sequential attempts to explain features, where categories are derived first, then features that are poorly explained are recategorized. We present two sets of simulations and experiments testing the models’ predictions about human categorization. We find that an approach based on joint inference provides the best fit to human categorization behavior, and we suggest that a full account of human category learning will need to incorporate something akin to these capabilities.

**LITERATURE SURVEY:**

Title : Learning Multiple Tasks with Kernel Methods

Author : TheodorosEvgeniou, Charles A. Micchelli, MassimilianoPontil

Year : 2013

**Description:**

We study the problem of learning many related tasks simultaneously using kernel methods and regularization. The standard single-task kernel methods, such as support vector machines and regularization networks, are extended to the case of multi-task learning. Our analysis shows that the problem of estimating many task functions with regularization can be cast as a single task learning problem if a family of multi-task kernel functions we define is used. These kernels model relations among the tasks and are derived from a novel form of regularizers. Specific kernels that can be used for multi-task learning are provided and experimentally tested on two real data sets. In agreement with past empirical work on multi-task learning, the experiments show that learning multiple related tasks simultaneously using the proposed approach can significantly outperform standard single-task learning particularly when there are many related tasks but few data per task.

**LITERATURE SURVEY:**

Title : A Framework for Learning Predictive Structures from

Multiple Tasks and Unlabeled Data

Author : Rie Kubota Ando, Tong Zhang tzhang.

Year : 2013

**Description:**

One of the most important issues in machine learning is whether one can improve the performance of a supervised learning algorithm by including unlabeled data. Methods that use both labeled and unlabeled data are generally referred to as semi-supervised learning. Although a number of such methods are proposed, at the current stage, we still don’t have a complete understanding of their effectiveness. This paper investigates a closely related problem, which leads to a novel approach to semi-supervised learning. Specifically we consider learning predictive structures on hypothesis spaces (that is, what kind of classifiers have good predictive power) from multiple learning tasks. We present a general framework in which the structural learning problem can be formulated and analyzed theoretically, and relate it to learning with unlabeled data. Under this framework, algorithms for structural learning will be proposed, and computational issues will be investigated. Experiments will be given to demonstrate the effectiveness of the proposed algorithms in the semi-supervised learning setting.

**LITERATURE SURVEY:**

Title : Multitask Learning

Author : Aarti Singh, ManishaMalhotra.

Year : 2013

**Description:**

Multitask Learning is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias. It does this by learning tasks in parallel while using a shared representation; what is learned for each task can help other tasks be learned better. This paper reviews prior work on MTL, presents new evidence that MTL in backprop nets discovers task relatedness without the need of supervisory signals, and presents new results for MTL with k-nearest neighbor and kernel regression. In this paper we demonstrate multitask learning in three domains. We explain how multitask learning works, and show that there are many opportunities for multitask learning in real domains. We present an algorithm and results for multitask learning with case-based methods like k-nearest neighbor and kernel regression, and sketch an algorithm for multitask learning in decision trees. Because multitask learning works, can be applied to many different kinds of domains, and can be used with different learning algorithms, we conjecture there will be many opportunities for its use on real-world problems.

**LITERATURE SURVEY:**

Title : Feature Selection for Ordinal Text Classification

Author : Stefano Baccianella, Andrea Esuli, and FabrizioSebastiani

Year : 2014

**Description:**

Ordinal classification (also known as ordinal regression) is a supervised learning task that consists of automatically determining the implied rating of a data item on a fixed, discrete rating scale. This problem is receiving increased attention from the sentiment analysis / opinion mining community, due to the importance of automatically rating increasing amounts of product review data in digital form. As in other supervised learning tasks such as (binary or multiclass) classification, feature selection is needed in order to improve efficiency and to avoid overfitting. However, while feature selection has been extensively studied for other classification tasks, is has not for ordinal classification. In this paper we present four novel feature selection metrics that we have specifically devised for ordinal classification, and test them on two datasets of product review data against three metrics previously known from the literature, using two learning algorithms from the “support vector regression” tradition. The experimental results show that all four proposed metrics largely outperform all of the three baseline techniques, on both datasets and for both learning algorithms.

**LITERATURE SURVEY:**

Title: Adaptive oating search methods in feature selection

Author : P. Somol, P. Pudil, J. Novovi\_cov\_a, P. Pacl\_õk.

Year : 2013

**Description:**

A new suboptimal search strategy for feature selection is presented. It represents a more sophisticated version of ``classical'' ¯oating search algorithms (Pudil et al., 1994), attempts to remove some of their potential de®ciencies and facilitates a solution even closer to the optimal one.

**LITERATURE SURVEY:**

Title : On p-norm Path Following in Multiple Kernel Learning for Non-linear

featureSelection

Author :Pratik Jawanpuria, ManikVarma, J. SakethaNath.

Year : 2009

**Description:**

Our objective is to develop formulations and algorithms for efficiently computing the feature selection path – i.e. the variation in classification accuracy as the fraction of selected features is varied from null to unity. We propose a novel conjecture which states that, for certain lp-MKL formulations, the number of features selected in the optimal solution monotonically decreases as p is decreased from an initial value to unity. We prove the conjecture, for a generic family of kernel target alignment based formulations, and show that the feature weights themselves decay (grow) monotonically once they are below (above) a certain threshold at optimality. This allows us to develop a path following algorithm that systematically generates optimal feature sets of decreasing size. The proposed algorithm sets certain feature weights directly to zero for potentially large intervals of p thereby reducing optimization costs while simultaneously providing approximation guarantees. We empirically demonstrate that our formulation can lead to classification accuracies which are as much as 10% higher on benchmark data sets not only as compared to other lp-MKL formulations and uniform kernel baselines but alsoleading feature selection methods. In particular, we generate the entire feature selection path for data sets with a hundred thousand features in approximately half an hour on standard hardware. Entire path generation for such data set is well beyond the scaling capabilities of other methods.

**MODULES:**

1**.**USER INTERFACE DESIGN:

2**.**FILE UPLOADING:

3.STORED IN DATABASE:

4. READ THE FILE AND UPLOADING FILE:

5. DEDUPLICATION:

**MODULE DESCRIPTION**

**1. User Interface Design:**

To connect with server user must give their username and password then only they can able to connect the server. If the user already exits directly can login into the server else user must register their details such as username, password and Email id, into the server. Server will create the account for the entire user to maintain upload and download rate. Name will be set as user id. . Logging in is usually used to enter a specific page.

**2. File Uploading**:

In this module the user logged in, then user are going to upload their file

3.**Store In Database**:

In this module, the user uploading all the file where stored in the database. The user retrieve the information from database.

4.**Read The File And Uploading File:**

In this module, he uploading file are stored in database and we are going to read the filename, filesize, filetype etc…, where all the information are stored in database.

**5.Deduplication:**

In this module, the user will uploading the file only once. Incase the user will uploading same file again will became as file already exist.

**MODULE DIAGRAM:**

1. User Interface Design:

Database

Welcome Page

Cloud Appplicatoin

Registration

Page

Login

Server

**2.FILE UPLOADING:**

**user login A.txt**

**3. STORE IN THE DATABASE :**

**User login A.txt fileDatabase**

**4. Read The File And Uploading File:**

**database**

**user login A.txt**

**A.txt**

**5.Deduplication:**

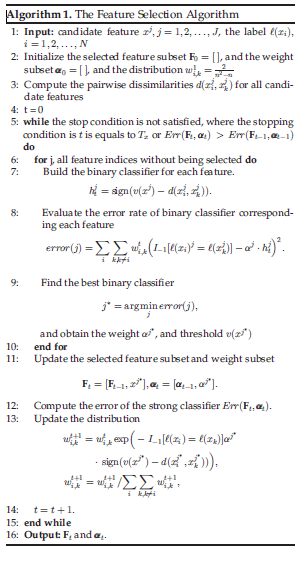
**Database**

**user login A.txt**

**A.txt**

**Duplicate file**

**SYSTEM ALGORITHM:**

****

**SYSTEM REQUIREMENTS**

**HARDWARE**

PROCESSOR : PENTIUM IV 2.60 GHz, Intel Dual Core.

RAM : 4 GB DD RAM

MONITOR : 15” LCD, LED MONITOR

HARD DISK : 40 GB

**SOFTWARE**

Front End : JAVA (j2ee, Servlets, Jsp)

Back End : My SQL

Operating System : Windows, Mac, Linux

IDE : Net Beans, Eclipse

Use Case Diagram:



**EXPLANATION:**

The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

**Class Diagram**:



**EXPLANATION:**

The class diagram is the main building block of object oriented modeling. It is used both for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code.

Object Diagram

a.txt database

user login file upload

a.txt

**EXPLANATION:**

Object diagram we are telling about the flow of objects how the process is running. In the above digram tells about the flow of objects between the classes.

‘

State Chart Diagram:



**EXPLANATION:**

State diagrams require that the system described is composed of a finite number of states; sometimes, this is indeed the case, while at other times this is a reasonable abstraction.

Activity Diagram:

user

login

File upload

a.txt

a.txt

Cloud

**EXPLANATION:**

In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

**Sequence Diagram**:



**EXPLANATION:**

In our sequence diagram specifying processes operate with one another and in order. In our sequence diagram first Data mining user login into Datamining.

**Collaboration Diagram**:



**EXPLANATION:**

A collaboration diagram describes interactions among objects in terms of sequenced messages. Collaboration diagrams represent a combination of information taken from class, sequence, and use case diagrams describing both the static structure and dynamic behavior of a system.

Data Flow Diagram:

Level 1:

Database

User Login

Level 2:

user

login

database

File uploading

File uploading

user

login

login

user

File uploading

login

login

login

user

user

login

File uploading

File uploading

Duplicate

not access

**EXPLANATION:**

It does not show information about the timing of processes, or information about whether processes will operate in sequence or in parallel. In the DFDs the level zero process is based on the login validations.

E-R Diagram:

USER

login

File 1

File 2

File 4

File3

**EXPLANATION:**

Entity-Relationship Model (ERM) is an abstract and conceptual representation of data. Entity-relationship modeling is a database modeling method, used to produce a type of conceptual schema or semantic data model of a system, often a relational database.

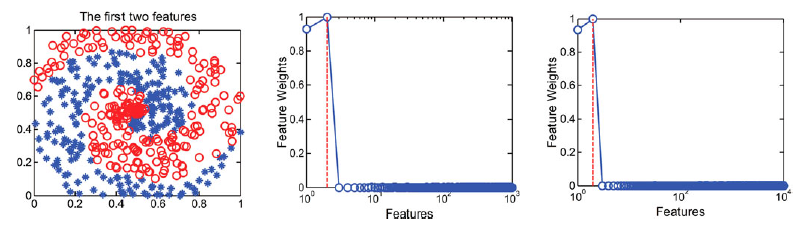
**Component Diagram**:



**EXPLANATION:**

In the Unified Modeling Language, a component diagram depicts how components are wired together to form larger components and they are used to illustrate the structure of arbitrarily complex systems.

**System Architecture**:



**FUTURE ENHANCEMENT**

Algorithm for frequent item set mining and association rule learning overtransactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

**ADVANTAGES:**

* Response time increased.
* Website user will increase.

**Conclusion:**

In this paper, a newly algorithm of feature evaluation is developed to measure the quality of feature, and applied to as a feature selection criterion. A feature subset that gives rise to higher classification ability is considered to be more important. With this criterion, the feature selection task is transformed into an optimization problem. The optimization problem is efficiently solved by following the principle of the AdaBoost-based search method, rather than the exhaustive search. In addition, we also analyze the generalization error bounds of our feature selection algorithm. Various experiments have been conducted on four UCI and 12 microarray data sets to demonstrate the effectiveness of our algorithm, and verify the theoretical results established in this paper.

**REFERENCES:**

[1] A. A. Alizadeh, et al., “Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling,” Nature, vol. 286, no. 5439, pp. 531–537, 1999.

[2] A. Arauzo-Azofra, J. Aznarte, and J. Benıtez, “Empirical study of feature selection methods based on individual feature evaluation for classification problems,” Expert Syst. Appl., vol. 38, no. 7, pp. 8170–8177, 2011.

[3] A. J. Stephenson, A. Smith, M. W. Kattan, J. Satagopan, V. E. Reuter, P. T. Scardino, and W. L. Gerald, “Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy,” Cancer, vol. 104, no. 2, pp. 290–298, 2005.

[4] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: A review,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, pp. 4–37, Jan. 2000.

[5] A. Tayal, T. Coleman, and Y. Li, “Primal explicit max margin feature selection for nonlinear support vector machines,” Pattern Recognit., vol. 47, no. 6, pp. 2153–2164, 2014.

[6] H. Almuallim and T. Dietterich, “Learning boolean concepts in the presence of many irrelevant features,” Artif.Intell., vol. 69, pp. 279–305, 1994.

[7] C. Ambroise and G. J McLachlan, “Selection bias in gene extraction on the basis of microarray gene-expression data,” Proc. Nat. Acad. Sci. USA, vol. 99, no. 10, pp. 6562–6566, 2002.

[8] C. Bishop, Neural Networks for Pattern Recognition. Oxford, U.K.: Oxford Univ. Press, 1995.

[9] C. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[10] V. Bolon-Canedo, N. Sanchez-Maro~no, and A. Alonso-Betanzos, “A review of feature selection methods on synthetic data,” Knowl.Inf. Syst., vol. 34, no. 3, pp. 483–519, 2013.