Study Notes for Comp 4350 Software Engineering II Lecture 8

AIOps (Artificial Intelligence for IT Operations)

- **Definition**: AIOps combines big data, machine learning (ML), and visualization to enhance IT operations.
- **Application**: Addresses DevOps challenges using AI.
- **Source**: Gartner, 2018.

Sub-topics of AIOps Research

- Automated logging
- Log abstraction
- Anomaly detection
- Performance analysis
- Incident prediction
- Logging practices
- Fault diagnostics
- System comprehension
- Monitoring code and data usage
- Autonomous configuration
- AIOps infrastructures

Agenda

- Hands-on tutorial for anomaly detection
- Usage of ML techniques in practice

Hands-on Tutorial for Anomaly Detection

- Reference: He et al., ISSRE '16.
- Process:
    1. Log Collection
    2. Log Parsing
    3. Feature Extraction
    4. Anomaly Detection
- Goal: Detect anomaly behavior in Hadoop DFS blocks.
- Resource: SEII-AIOps_data GitHub repository

Follow-up Tasks for Anomaly Detection

- Correlation analysis and PCA for dimension reduction.
- Mining frequent patterns in logs.
- Clustering log sequences based on log event counts.
- Classifying anomalies using a bag of words, treating log level as ordinal.

Common Machine Learning Techniques

- Classification: Organizing data into predefined categories.
- Clustering: Grouping similar data.
- Dimension reduction: Simplifying data without losing significant information.

Classification Process

- Steps: Training/testing sets, model construction, evaluation.
- Evaluation metrics: Accuracy, Precision, Recall, F-measure, AUC, MAP.

Machine Learning Data Types

- Categories: Categorical, Numerical, Nominal, Binary, Ordinal, Continuous, Discrete.
- Examples: True/False, language names, skill level, height, weight range.

Encoding Categorical Data

- Converting categorical data into numerical for ML processing.

Ordinal Data in Logging

- Metrics: Logging statement, block, file, code change, historical change.
- Levels: Trace, Debug, Info, Warn, Error, Fatal.
- Use: Ordinal Regression Model (Li et al., EMSE '17).

Handling Different Data Types

- Choose the appropriate model based on data type compatibility.
- Models: Random forest, logistic regression, support vector machine, ordinal regression.

Clustering Techniques

- Hierarchical and Partitional.
- K-means algorithm steps: Decide on k value, initialize centers, assign classes, re-estimate centers, iterate until no change.

Similarity Measures

- Introduction: Similarity is fundamental in ML techniques.
- Measures: Euclidean distance, cosine similarity, correlation distance, Jaccard similarity.

Performance Regression Analysis

- Analyzing perf metrics (CPU, memory, response time) alongside logs.

High Dimensionality and Configuration Parameters

- Exhaustive testing of configurations is infeasible due to combinatorial explosion.
- Use domain experts and tests to identify performance critical parameters.

- Explore Multivariate Adaptive Regression Splines (MARS).

Feature Importance Techniques

- Evaluate how much a feature contributes to the prediction power of a model.
- Permutation feature importance - model-agnostic approach.

Dimension Reduction Techniques

- PCA: Transforming correlated variables into uncorrelated principal components.
- Correlation analysis: Identifying and eliminating redundant variables.
- t-SNE: Dimensionality reduction for visualization of high-dimensional datasets.

Scenario-Based Technique Selection

- Example: Select techniques for analyzing user behavior in an e-commerce system.
- Options: Classification, Clustering, Frequent Pattern Mining, Dimension Reduction.

Tools and Datasets

- Tools: APM tools (e.g. AppDynamics, New Relic), Log Management tools (e.g. Splunk, ELK Stack).
- Academic tools: LogPAI.
- Datasets: Loghub, Azure VM traces, Google cluster traces, Alibaba cluster traces.

Reference Material

- Notable papers and research in the domain of AIOps.

Class Notes

Slide 1

Title: Comp 4350 Software Engineering II Lecture 8
Dr. Shaowei Wang

Slide 2

Title: AIOps (Artificial Intelligence for IT Operations) AIOps enhances IT operations through greater insights by combining big data, machine learning and visualization. AIOps addresses the DevOps challenges with AI. From: 2018 Gartner

Slide 3

Title: Sub-topics of AIOps research

- Automated logging
- Log abstraction
- Anomaly detection

- Performance analysis
- Incident prediction
- Logging practices
- Fault diagnostics
- System comprehension
- Monitoring code made right
- Monitoring data used right
- Autonomous configuration
- AIOps infrastructures

Slide 4

Title: Agenda

- Hand-on tutorial for anomaly detection
- Practices of using machine learning techniques

Slide 5

Title: Hand-on tutorial for Anomaly Detection

Slide 6

Title: Hands-on tutorial: Analyzing System Logs for Anomaly Detection He et al., Experience Report: System Log Analysis for Anomaly Detection. ISSRE ‘16.

- Log Collection
- Log Parsing
- Feature Extraction
- Anomaly Detection
- Hadoop distributed file system(DFS)

Slide 7

Title: Hands-on tutorial: Analyzing System Logs for Anomaly Detection He et al., Experience Report: System Log Analysis for Anomaly Detection. ISSRE ‘16.

- Log Collection
- Log Parsing
- Feature Extraction
- Anomaly Detection Goal: Detect anomaly behavior on each block

Slide 8

Title: Hands-on tutorial: Analyzing System Logs for Anomaly Detection He et al., Experience Report: System Log Analysis for Anomaly Detection. ISSRE ‘16.

- Log Collection
- Log Parsing
- Feature Extraction

- Anomaly Detection GitHub repo: SEII-AIOps_data

Slide 9

Title: Hands-on tutorial: follow-up tasks

- Using correlation analysis to reduce the feature dimension before fitting the model
- Using PCA analysis to reduce the feature dimension before fitting the model
- Mining frequent patterns in the logs and summarize the frequent normal and abnormal patterns
- Clustering log sequences based on the counts of each log event in a sequence
- Classify the anomaly of each log line using a bag of words and treat log level as an ordinal variable

Slide 10

Title: Common machine learning techniques

- Classification
- Clustering
- Dimension reduction

Slide 11

Title: Two steps process - classification

- Training and Testing sets, Model construction, Model Evaluation
- Evaluation metrics: Accuracy, Precision, Recall, F-measure, AUC, MAP, etc.

Slide 12

Title: Model Application From: A. E. Hassan and T. Xie: Mining Software Engineering Data

Slide 13

Title: Different types of data in supervised learning

- Data Types: Categorical, Numerical, Nominal, Binary, Ordinal, Continuous, Discrete

Slide 14

Title: Encoding categorical data into numerical data

- Data Types: Categorical, Numerical, Nominal, Binary, Ordinal, Continuous, Discrete

Slide 15

Title: Ordinal data

- Logging statement metrics
- Containing block metrics
- Containing file metrics
- Code change metrics
- Historical change metrics

- Trace, Debug, Info, Warn, Error, Fatal
- Ordinal Regression Model Li et al., Which log level should developers choose for a new logging statement? EMSE ‘17.

Slide 16

Title: Handling different types of data in supervised learning First of all, consider the right models Some models cannot handle categorical data directly (e.g., linear regression model) Some models can handle categorical response variables but not categorical explanatory variables (e.g., logistic regression model) Some models can handle both categorical response variables and categorical explanatory variables (e.g., decision tree-based models, random forest) Some models provide special handling for certain data types (e.g., ordinal regression model for ordinal data)

Slide 17

Title: Commonly used classification models

- Random forest: construct a forest of decision trees, and get the final prediction based on the votes of each decision tree.

Slide 18

Title: Commonly used classification models

- Logistic regression

Slide 19

Title: Commonly used classification models

- Support vector machine: Map the data points into a space, so that the margin between two classes is maximized.

Slide 20

Title: Starting from the simple one first

- Logistic regression, which is explainable: understanding indicator is the most important one to predict a failure of cloud node
- Random forest: Achieve robust performance

Slide 21

Title: Common machine learning techniques

- Classification
- Clustering
- Dimension reduction

Slide 22

Title: Two types of clustering

- Hierarchical clustering
- Partitional clustering

Slide 23

Title: Algorithm k-means

1. Decide on a value for k (number of clusters).
2. Initialize the k cluster centers (randomly, if necessary).
3. Decide the class memberships of the N objects by assigning them to the nearest cluster center.
4. Re-estimate the k cluster centers, by assuming the memberships found above are correct.
5. If none of the N objects changed membership in the last iteration, exit. Otherwise go to 3.

Slide 24

Title: The K-Means Clustering Method K=2

- Arbitrarily partition the objects into K clusters
- Compute the cluster means
- Update the cluster means
- Reassign

Slide 25

Title: Similarity is widely used in various machine learning techniques.

Slide 26

Title: Clustering: the process of grouping a set of objects into classes of similar objects

Slide 27

Title: Anomaly detection

Slide 28

Title: But, what is similarity? The quality or state of being similar; likeness; resemblance; as, a similarity of features. Webster's Dictionary

Slide 29

Title: Feature engineering

Slide 30

Define cost of alignment (edit cost) and get optimal alignment.

Slide 31

Title: Similarity measures How to determine similarity between data points using various distance metrics Euclidean distance

Slide 32

Title: Cosine similarity

Slide 33

Title: Similarity measures Correlation distance

Slide 34

Title: Jaccard Similarity

Slide 35

Title: Common machine learning techniques

- Classification
- Clustering
- Dimension reduction

Slide 36

Title: Perf regression analysis usually needs to examine hundreds of perf metrics

- Perf metrics: CPU, Memory, Response time
- Logs: Warnings, Errors, Exceptions
- Heavy tasks for performance engineers

Slide 37

Title: Blue force approach One way to fully understand the effect of the configuration parameters (more than 100) on system performance is to exhaustively run different combinations of the parameter values. However, too many combinations to test become infeasible! Each combination requires several hours to run. What if coming with hundreds of parameters?

Slide 38

Title: Dealing with high dimensionality

Slide 39

Title: Understanding the relationship between config. parameters and performance metrics Asking domain experts

- Perf. related parameters
- Perf. data
- Perf. critical parameters
- Running tests
- Only a few out of many candidate parameters significantly impact system performance and reduce the search space.
- Multivariate Adaptive Regression Splines (MARS) [Li et al. 2018]

Slide 40

Title: Feature importance How much contribution of a feature (e.g., Cache.size, Fetch.blocksize) makes to the prediction power of a model (e.g., average query response time). More contribution of a feature makes, more important it is

Slide 41

Title: Permutation feature importance After evaluating the performance of your model, you permute the values of a feature of interest and reevaluate model performance. If more performance drops due to the permutation, more importance it is; otherwise, less importance. It is a model-agnostic approach

Slide 42

Title: Dimension reduction – Principle Component Analysis (PCA) PCA transforms a set of correlated variables into a reduced set of uncorrelated variables (principle components, or PC). Each PC is a linear combination of the original variables. A smaller number of PCs can explain the variance of the original variables

Slide 43

Title: Dimension reduction – correlation analysis Y = a1*X1 + a2 * X2 + a3 * X3. If X1 = b2 * X2 + b3 * X3, we can remove X1, and since we can use X2 and X3 to represent X1.

Slide 44

Title: Dimension reduction – correlation analysis

Slide 45

Title: Other dimension reduction techniques

- t-Distributed Stochastic Neighbor Embedding (t-SNE): A prize-winning technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. Idea is simple: models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

Slide 46

Title: Exercise: select the most appropriate techniques for different scenarios Your monitoring data recorded user behaviors (e.g., browsing items, purchasing items, etc.) in an E-commerce system. You want to find the patterns of user behaviors (e.g.., browsing item A following purchasing item B). Which technique will you use?

- Classification
- Clustering
- Frequent Pattern Mining
- Dimension Reduction

Slide 47

Title: Tools and Datasets

Slide 48

Title: Available tools

- Industrial solutions:
  - Application Performance Management (APM) Tools: e.g., AppDynamics, New Relic, Dynatrace, and Pinpoint (open source)
  - Log Management Tools: e.g., Splunk, ELK Stack (open source), Loggly, and Graylog
- Academic tools:
  - LogPAI (https://github.com/logpai) : log parsers, anomaly detection, where to log

Slide 49

Title: Public Datasets

- Loghub (https://github.com/logpai/loghub): a collection of system logs generated by different systems
- Azure VM traces (https://github.com/Azure/AzurePublicDataset): two representative traces of the virtual machine (VM) workload of Microsoft Azure collected in 2017 and 2019.
- Google cluster traces (https://github.com/google/cluster-data/blob/master/ClusterData2011_2.md)
- Alibaba cluster traces (https://github.com/alibaba/clusterdata/wiki/About-Alibaba-cluster-and-why-we-open-the-data)

Slide 50

Title: Reference Li, Yangguang, et al. "Predicting Node Failures in an Ultra-large-scale Cloud Computing Platform: an AIOps Solution." TOSEM ‘20.
Lin et al., Predicting Node Failure in Cloud Service Systems. FSE ‘18.
Zhu et a., Learning to log: helping developers make informed logging decisions. ICSE ‘15.
El-Sayed et al. "Learning from failure across multiple clusters: A trace-driven approach to understanding, predicting, and mitigating job terminations." ICDCS ‘17.
Botezatu et al. "Predicting disk replacement towards reliable data centers." SIGKDD ‘16.
Xu, Wei, et al. "Detecting large-scale system problems by mining console logs." SOSP ‘09.