ELYSIUMPRO

A UNIT OF ELYSIUM GROUPS

FINAL YEAR PROJECT

DATA MINING 2019-2020

TITLES WITH ABSTRACTS

CALL US @

(+91) 9944 7933 98 | (+91) 452 - 424 2842, 424 2843

20 Years of Experience | Automated Services | 24/7 Help Desk Support
Advanced Technologies and Tools | Legitimate Members of all Journals
Quality Project Training | Industry Exposure

ELYSIUMPRO

INSPIRING THE LEADING EDGE TECHNOLOGIES

www.elysiumpro.in

#227, Elysium Campus, Church Road, Anna Nagar,
Madurai - 625020, Tamil Nadu, India

CALL US @

(+91) 9944 7933 98 | (+91) 452 - 424 2842, 424 2843

/ElysiumPro Project Center / ElysiumPro / ElysiumPro

ELYSIUM GROUPS

EPRO DM - 001	Privacy-Preserving User Profile Matching in Social Networks

In this paper, we consider a scenario where a user queries a user profile database, maintained by a social networking service provider, to identify users whose profiles match the profile specified by the querying user. A typical example of this application is online dating. Most recently, an online dating website, Ashley Madison, was hacked, which results in disclosure of a large number of dating user profiles. This data breach has urged researchers to explore practical privacy protection for user profiles in a social network. In this paper, we propose a privacy-preserving solution for profile matching in social networks by using multiple servers. Our solution is built on homomorphic encryption and allows a user to find out matching users with the help of multiple servers without revealing to anyone the query and the queried user profiles in clear. Our solution achieves user profile privacy and user query privacy as long as at least one of the multiple servers is honest. Our experiments demonstrate that our solution is practical.

EPRO DM - 002	Graph K-means based on Leader Identification, Dynamic Game and Opinion Dynamics

With the explosion of social media networks, many modern applications are concerning about people's connections, which leads to the so-called social computing. An elusive question is to study how opinion communities form and evolve in real-world networks with great individual diversity and complex human connections. In this paper, we attempt to model a realistic social media network as a discrete-time dynamical system, where the opinion matrix and the community structure could mutually affect each other. The community detection in social media networks is naturally formulated as a multi-objective optimization problem, i.e., finding a set of densely connected components with similar opinion vectors. We propose a novel and powerful graph K-means framework, which is composed of three coupled phases in each discrete-time period. Specifically, the first phase uses a fast heuristic approach to identify those opinion leaders who have relatively high local reputation; the second phase adopts a novel dynamic game model to find the locally Pareto-optimal community structure; the final phase employs a robust opinion dynamics model to simulate the evolution of the opinion matrix. We conduct a series of comprehensive experiments on real-world benchmark networks to validate the performance of GK-means through comparisons with the state-of-the-art graph clustering technologies.

EPRO DM - 003	Multi-view Scaling Support Vector Machines for Classification and Feature Selection

With explosive growth of data, the multi-view data is widely used in many fields, such as data mining, machine learning, computer vision and so on. Because such data always has complex structure, i.e. many categories, many perspectives of description and high dimension, how to formulate an accurate and reliable framework for multi-view classification is a very challenging task. In this paper, we propose a novel multi-view classification method by using multiple multi-class support vector machines (SVMs) and a novel collaborative strategy. Here each multi-class SVM integrates the scaling factor to renewedly adjust the weight allocation which is beneficial to highlight some more discriminative features. Furthermore, we use the decision function values of multiple learners to combine multiple multi-class learners, and then determine the final classification results according to a final confidence score. In addition, through a series of theoretical analyses, we bridge the proposed model with the solvable problem and solve it by an iterative optimization method with convergence. We evaluate the proposed method on several image datasets and face datasets, and the experimental results demonstrate that our proposed method performs better than other state-of-the-art learning algorithms.

EPRO DM - 004	Mining Behavioral Sequence Constraints for Classification

Sequence classification deals with the task of finding discriminative and concise sequential patterns. To this purpose, many techniques have been proposed, which mainly resort to the use of partial orders to capture the underlying sequences in a database according to the labels. Partial orders, however, pose many limitations, especially on expressiveness, i.e. the aptitude towards capturing certain behavior, and on conciseness, i.e. doing so in a compact and informative way. These limitations can be addressed by using a better representation. In this paper we present the interesting Behavioral Constraint Miner (iBCM), a sequence classification technique that discovers patterns using behavioral constraint templates. The templates comprise a variety of constraints and can express patterns ranging from simple occurrence, to looping and position-based behavior over a sequence. Furthermore, iBCM also captures negative constraints, i.e. absence of particular behavior. The constraints can be discovered by using simple string operations in an efficient way. Finally, deriving the constraints with a window-based approach allows to pinpoint where the constraints hold in a string, and to detect whether patterns are subject to concept drift. Through empirical evaluation, it is shown that iBCM is better capable of classifying sequences more accurately and concisely in a scalable manner.

EPRO DM - 005	A framework for supervised classification performance analysis with information-theoretic methods

We introduce a framework for the evaluation of multiclass classifiers by exploring their confusion matrices. Instead of using error-counting measures of performance, we concentrate in quantifying the information transfer from true to estimated labels using information-theoretic measures. First, the Entropy Triangle allows us to visualize the balance of mutual information, variation of information and the deviation from uniformity in the true and estimated label distributions. Next the Entropy-Modified Accuracy allows us to rank classifiers by performance while the Normalized Information Transfer rate allows us to evaluate classifiers by the amount of information accrued during learning. Finally, if the question rises to elucidate which errors are systematically committed by the classifier, we use a generalization of Formal Concept Analysis to elicit such knowledge. All such techniques can be applied either to artificially or biologically embodied classifiers---e.g. human performance on perceptual tasks. We instantiate the framework in a number of examples to provide guidelines for the use of these tools in the case of assessing single classifiers or populations of them---whether induced with the same technique or not---either on single tasks or in a set of them. These include UCI tasks and the more complex KDD cup 99 competition on Intrusion Detection.

EPRO DM - 006	GERF: a group event recommendation framework based on learning-to-rank

Event recommendation is an essential means to enable people to find attractive upcoming social events, such as party, exhibition and concert. While growing line of research has focused on suggesting events to individuals, making event recommendation for a group of users has not been well studied. In this paper, we aim to recommend upcoming events for a group of users. We formalize group recommendation as a ranking problem and propose a group event recommendation framework GERF based on learning-to-rank technique. Specifically, we first analyze different contextual influences on user's event attendance, and extract preference of user to event considering each contextual influence. Then, the preference scores of the users in a group are taken as the features for learning-to-rank to model the preference of the group. Moreover, a fast pairwise learning-to-rank algorithm, Bayesian group ranking, is proposed to learn ranking model for each group. Our framework is easily to incorporate additional contextual influences, and can be applied to other group recommendation scenarios. Extensive experiments have been conducted to evaluate the performance of GERF on two real-world datasets and demonstrate the appealing performance of our method on both accuracy and time efficiency.

EPRO DM - 007	A Joint Two-Phase Time-Sensitive Regularized Collaborative Ranking Model for Point of Interest Recommendation

The popularity of location-based social networks (LBSNs) has led to a tremendous amount of user check-in data. Recommending points of interest (POIs) plays a key role in satisfying users needs in LBSNs. While recent work has explored the idea of adopting collaborative ranking (CR) for recommendation, there have been few attempts to incorporate temporal information for POI recommendation using CR. In this article, we propose a two-phase CR algorithm that incorporates the geographical influence of POIs and is regularized based on the variance of POIs popularity and users activities over time. The time-sensitive regularizer penalizes user and POIs that have been more time-sensitive in the past, helping the model to account for their long-term behavioral patterns while learning from user-POI interactions. Moreover, in the first phase, it attempts to rank visited POIs higher than the unvisited ones, and at the same time, apply the geographical influence. In the second phase, our algorithm tries to rank users favorite POIs higher on the recommendation list. Both phases employ a collaborative learning strategy that enables the model to capture complex latent associations from two different perspectives. Experiments on real-world datasets show that our proposed time-sensitive collaborative ranking model beats state-of-the-art POI recommendation methods.

EPRO DM - 008	Approximate Nearest Neighbor Search on High Dimensional Data - Experiments, Analyses, and Improvement

Approximate Nearest neighbor search (ANNS) is fundamental and essential operation in applications from many domains, such as databases, machine learning, multimedia, and computer vision. Although many algorithms have been continuously proposed in the literature in the above domains each year, there is no comprehensive evaluation and analysis of their performances. In this paper, we conduct a comprehensive experimental evaluation of many state-of-the-art methods for approximate nearest neighbor search. Our study (1) is cross-disciplinary (i.e., including 16 algorithms in different domains, and from practitioners) and (2) has evaluated a diverse range of settings, including 20 datasets, several evaluation metrics, and different query workloads. The experimental results are carefully reported and analyzed to understand the performance results. Furthermore, we propose a new method that achieves both high query efficiency and high recall empirically on majority of the datasets under a wide range of settings

EPRO DM - 009	Similarity Search for Dynamic Data Streams

Nearest-neighbor searching systems are an integral part of many online applications, including but not limited to pattern recognition, plagiarism detection and recommender systems. With increasingly larger data sets, scalability has become an important issue. Many of the most space and running time efficient algorithms are based on locality sensitive hashing. The de facto standard approach to quickly answer nearest-neighbor queries on such a data set is usually a form of min-hashing. Not only is min-hashing very fast, but it is also space efficient and can be implemented in many computational models aimed at dealing with large data sets such as MapReduce and streaming. A significant drawback is that minhashing and related methods are only able to handle insertions to user profiles and tend to perform poorly when items may be removed. We initiate the study of scalable locality sensitive hashing (LSH) for dynamic data-streams. Specifically, using the Jaccard index as similarity measure, we design (1) a nearest-neighbor datastructure maintainable in dynamic data streams and (2) a sketching algorithm for similarity estimation. Our algorithms have little overhead in terms of running time compared to previous LSH approaches for the insertion streams, and drastically outperform previous algorithms in case of deletions

EPRO DM - 010	Practical Multi-Keyword and Boolean Search over Encrypted E-mail in Cloud Server

With the outbreak of e-mail message leakage events, such as the Hillary Clinton's Email Controversy, privacy and security of sensitive e-mail information have become users' primary concern. Encrypted email seems to be a viable solution for providing security, but it will greatly limit their operations. Public encryption with keyword search (PEKS) scheme is a popular technology to incorporate security protection and favorable operability functions together, which can play an important role in searching over encrypted email in a cloud server. In this paper, we propose a practical PEKS scheme named as public-key multi-keyword searchable encryption with hidden structures (PMSEHS). It could enable e-mail receivers to do the multi-keyword and boolean search in the large encrypted email database as fast as possible, without revealing more information to the cloud server. We also give comparative experiments, which demonstrate that our scheme has a higher efficiency in multi-keyword search for encrypted emails.

EPRO DM - 011	Efficient Mining of Frequent Patterns on Uncertain Graphs

Uncertainty is intrinsic to a wide spectrum of real-life applications, which inevitably applies to graph data. Representative uncertain graphs are seen in bio-informatics, social networks, etc. This paper motivates the problem of frequent subgraph mining on single uncertain graphs, and investigates two different - probabilistic and expected - semantics in terms of support definitions. First, we present an enumeration-evaluation algorithm to solve the problem under probabilistic semantics. By showing the support computation under probabilistic semantics is #P-complete, we develop an approximation algorithm with accuracy guarantee for efficient problem-solving. To enhance the solution, we devise computation sharing techniques to achieve better mining performance. Afterwards, the algorithm is extended in a similar flavor to handle the problem under expected semantics, where checkpoint-based pruning and validation techniques are integrated. Experiment results on real-life datasets confirm the practical usability of the mining algorithms.

EPRO DM – 012	Variable Weighting in Fuzzy k-Means Clustering to Determine the Number of Clusters

One of the most significant problems in cluster analysis is to determine the number of clusters in unlabeled data, which is the input for most clustering algorithms. Some methods have been developed to address this problem. However, little attention has been paid on algorithms that are insensitive to the initialization of cluster centers and utilize variable weights to recover the number of clusters. To fill this gap, we extend the standard fuzzy k-means clustering algorithm. It can automatically determine the number of clusters by iteratively calculating the weights of all variables and the membership value of each object in all clusters. Two new steps are added to the fuzzy k-means clustering process. One of them is to introduce a penalty term to make the clustering process insensitive to the initial cluster centers.The other one is to utilize a formula for iterative updating of variable weights in each cluster based on the current partition of data. Experimental results on real-world and synthetic datasets have shown that the proposed algorithm effectively determined the correct number of clusters while initializing the different number of cluster centroids. We also tested the proposed algorithm on gene data to determine a subset of important genes.

EPRO DM - 013	Graph K-means based on Leader Identification, Dynamic Game and Opinion Dynamics

With the explosion of social media networks, many modern applications are concerning about people's connections, which leads to the so-called social computing. An elusive question is to study how opinion communities form and evolve in real-world networks with great individual diversity and complex human connections. In this paper, we attempt to model a realistic social media network as a discrete-time dynamical system, where the opinion matrix and the community structure could mutually affect each other. The community detection in social media networks is naturally formulated as a multi-objective optimization problem, i.e., finding a set of densely connected components with similar opinion vectors. We propose a novel and powerful graph K-means framework, which is composed of three coupled phases in each discrete-time period. Specifically, the first phase uses a fast heuristic approach to identify those opinion leaders who have relatively high local reputation; the second phase adopts a novel dynamic game model to find the locally Pareto-optimal community structure; the final phase employs a robust opinion dynamics model to simulate the evolution of the opinion matrix. We conduct a series of comprehensive experiments on real-world benchmark networks to validate the performance of GK-means through comparisons with the state-of-the-art graph clustering technologies.

EPRO DM - 014	GMC: Graph-based Multi-view Clustering

Multi-view graph-based clustering aims to provide clustering solutions to multi-view data. However, most existing methods do not give sufficient consideration to weights of different views and require an additional clustering step to produce the final clusters. They also usually optimize their objectives based on fixed graph similarity matrices of all views. In this paper, we propose a general Graph-based Multi-view Clustering (GMC) to tackle these problems. GMC takes the data graph matrices of all views and fuses them to generate a unified matrix. The unified matrix in turn improves the data graph matrix of each view, and also gives the final clusters directly. The key novelty of GMC is its learning method, which can help the learning of each view graph matrix and the learning of the unified matrix in a mutual reinforcement manner. A novel multi-view fusion technique can automatically weight each data graph matrix to derive the unified matrix. A rank constraint without introducing a tuning parameter is also imposed on the Laplacian matrix of the unified matrix, which helps partition the data points naturally into the required number of clusters. An alternating iterative optimization algorithm is presented to optimize the objective function. Experimental results demonstrate that the proposed method outperforms state-of-the-art baselines markedly.

EPRO DM - 015	Adaptive Self-paced Deep Clustering with Data Augmentation

Deep clustering gains superior performance than conventional clustering by jointly performing feature learning and cluster assignment. Although numerous deep clustering algorithms have emerged in various applications, most of them fail to learn robust cluster-oriented features which in turn hurts the final clustering performance. To solve this problem, we propose a two-stage deep clustering algorithm by incorporating data augmentation and self-paced learning. Specifically, in the first stage, we learn robust features by training an autoencoder with examples that are augmented by random shifting and rotating the given clean examples. Then in the second stage, we encourage the learned features to be cluster-oriented by alternatively finetuning the encoder with the augmented examples and updating the cluster assignments of the clean examples. During finetuning the encoder, the target of each augmented example in the loss function is the center of the cluster to which the clean example is assigned. The targets may be computed incorrectly, and the examples with incorrect targets could mislead the encoder network. To stabilize the network training, we select most confident examples in each iteration by utilizing the adaptive self-paced learning. Extensive experiments validate that our algorithm outperforms the state of the arts on four image datasets.

EPRO DM - 016	An Effective Clustering Method over CF+ Tree Using Multiple Range Queries

Many existing clustering methods usually compute clusters from the reduced data sets obtained by summarizing the original very large data sets. BIRCH is a popular summary-based clustering method that first builds a CF tree, and then performs a global clustering using the leaf entries of the tree. However, to the best of our knowledge, no prior studies have proposed a global clustering method that uses the structure of a CF tree. Therefore, we propose a novel global clustering method ERC(effective multiple range queries-based clustering), which takes advantage of the structure of a CF tree. We further propose a CF+ tree, which optimizes the node split scheme used in the CF tree. As a result, the CF+ ERC (CF+ tree-based ERC) method effectively computes clusters over large data sets. Furthermore, it does not require a predefined number of clusters to compute the clusters. We present in-depth theoretical and experimental analyses of our method. Experimental results on very large synthetic data sets demonstrate that the proposed approach is effective in terms of cluster quality and robustness and is significantly faster than existing clustering methods. In addition, we apply our clustering method to real data sets and achieve promising results.

EPRO DM - 017	ASCENT: Active Supervision for Semi-supervised Learning

Active learning algorithms attempt to overcome the labeling bottleneck by asking queries from large collection of unlabeled examples. Existing batch mode active learning algorithms suffer from three limitations: (1) The methods that are based on similarity function or optimizing certain diversity measurement, in which may lead to suboptimal performance and produce the selected set with redundant examples; (2) The models with assumption on data are hard in finding images that are both informative and representative; (3) The problem of noise labels has been an obstacle for algorithms. In this paper, we propose a novel active learning method that makes embeddings of labeled examples to those of unlabeled ones and back via deep neural networks. The active scheme makes correct association cycles that end up at the same class from that the association was started, which considers both the informativeness and representativeness of examples, as well as being robust to noise labels. We apply our active learning method to semi-supervised classification and clustering. The submodular function is designed to reduce the redundancy of the selected examples. Specifically, we incorporate our batch mode active scheme into the classification approaches, in which the generalization ability is improved. For semi-supervised clustering, we try to use our active scheme for constraints to make fast convergence and perform better than unsupervised clustering. Finally, we apply our active learning method to data filtering. To validate the effectiveness of the proposed algorithms, extensive experiments are conducted on diversity benchmark datasets for different tasks,

EPRO DM - 018	Cleaning Data with Forbidden Itemsets

Methods for cleaning dirty data typically employ additional information about the data, such as user-provided constraints specifying when data is dirty, e.g., domain restrictions, illegal value combinations, or logical rules. However, real-world scenarios usually only have dirty data available, without known constraints. In such settings, constraints are automatically discovered on dirty data and discovered constraints are used to detect and repair errors. Typical repairing processes stop there. Yet when constraint discovery algorithms are re-run on the repaired data (assumed to be clean), new constraints and thus errors are often found. The repairing process thus introduces new constraint violations. We present a different type of repairing method, which prevents introducing new constraint violations, according to a discovery algorithm. Summarily, our repairs guarantee that all errors identified by constraints discovered on the dirty data are fixed; and the constraint discovery process cannot identify new constraint violations. We do this for a new kind of constraints, called forbidden itemsets (FBIs), capturing unlikely value co-occurrences. We show that FBIs detect errors with high precision. Evaluation on real-world data shows that our repair method obtains high-quality repairs without introducing new FBIs. Optional user interaction is readily integrated, with users deciding how much effort to invest.

EPRO DM - 019	Social-aware Sequential Modeling of User Interests: A Deep Learning Approach

In this paper, we propose to leverage the emerging deep learning techniques for sequential modeling of user interests based on big social data, which takes into account influence of their social circles. First, we present a preliminary analysis for two popular big datasets from Yelp and Epinions. We show statistically sequential actions of all users and their friends, and discover both temporal autocorrelation and social influence on decision making, which motivates our design. Then, we present a novel hybrid deep learning model, Social-Aware Long Short-Term Memory (SA-LSTM), for predicting the types of item/PoIs that a user will likely buy/visit next, which features stacked LSTMs for sequential modeling and an autoencoder-based deep model for social influence modeling. Moreover, we show that SA-LSTM supports end-to-end training. We conducted extensive experiments for performance evaluation using the two real datasets from Yelp and Epinions. The experimental results show that (1) the proposed deep model significantly improves prediction accuracy compared to widely used baseline methods; (2) the proposed social influence model works effectively; and (3) going deep does help improve prediction accuracy but a not-so-deep deep structure leads to the best performance.

EPRO DM – 020	Mining Top-k Useful Negative Sequential Patterns via Learning

As an important tool for behavior informatics, negative sequential patterns (NSPs) (such as missing a medical treatment) are sometimes much more informative than positive sequential patterns (PSPs) (e.g., attending a medical treatment) in many applications. However, NSP mining is at an early stage and faces many challenging problems, including 1) how to mine an expected number of NSPs; 2) how to select useful NSPs; and 3) how to reduce high time consumption. To solve the first problem, we propose an algorithm Topk-NSP to mine the k most frequent negative patterns. In Topk-NSP, we first mine the top-k PSPs using the existing methods, and then we use an idea which is similar to top-k PSPs mining to mine the top-k NSPs from these PSPs. To solve the remaining two problems, we propose three optimization strategies for Topk-NSP. The first optimization strategy is that, in order to consider the influence of PSPs when selecting useful top-k NSPs, we introduce two weights, wP and wN, to express the user preference degree for NSPs and PSPs, respectively, and select useful NSPs by a weighted support wsup. The second optimization strategy is to merge wsup and an interestingness metric to select more useful NSPs. The third optimization strategy is to introduce a pruning strategy to reduce the high computational costs of Topk-NSP. Finally, we propose an optimization algorithm Topk-NSP$^+$. To the best of our knowledge, Topk-NSP$^+$ is the first algorithm that can mine the top-k useful NSPs. The experimental results on four synthetic and two real-life data sets show that the Topk-NSP$^+$ is very efficient in mining the top-k NSPs in the sense of computational cost and scalability.

EPRO DM - 021	Top-k Dominating Queries on Skyline Groups

The top- *k* dominating (TKD) query on skyline groups returns *k* skyline groups that dominate the maximum number of points in a given data set. The TKD query combines the advantages of skyline groups and top- *k* dominating queries, thus has been frequently used in decision making, recommendation systems, and quantitative economics. Traditional skylines are inadequate to answer queries from both individual and groups of points. The group size could be too large to be processed in a reasonable time as a single operator (i.e., the skyline group operator). In this paper, we address the performance problem of grouping for TKD queries in skyline database. We formulate the problem of grouping, define the group operator in skyline, and propose several efficient algorithms to find top- *k* skyline groups. Thus, we provide a systematic study of TKD queries on skyline groups and validate our algorithms with extensive empirical results on synthetic and realworld data

EPRO DM - 022	Answering Top-k Graph Similarity Queries in Graph Databases

Searching similar graphs in graph databases for a query graph has attracted extensive attention recently. Existing works on graph similarity queries are threshold based approaches which return graphs with distances to the query smaller than a given threshold. However, in many applications the number of answer graphs for the same threshold can vary significantly for different queries. In this paper, we study the problem of finding top-k most similar graphs for a query under the distance measure based on maximum common subgraph (MCS). Since computing MCS is NP-hard, we devise a novel framework to prune unqualified graphs based on the lower bounds of graph distance, and accordingly derive four lower bounds with different tightness and computational cost for pruning. To further reduce the number of MCS computations, we also propose an improved framework based on both lower and upper bounds, and derive three new upper bounds. To support efficient pruning, we design three indexes with different tradeoffs between pruning power and construction cost. To accelerate the index construction, we explore bound relaxation techniques, based on which approximate indexes can be efficiently built. We conducted extensive performance studies on large real datasets to validate the effectiveness and efficiency of our approaches.

EPRO DM - 023	Similarity Search for Dynamic Data Streams

Nearest-neighbor searching systems are an integral part of many online applications, including but not limited to pattern recognition, plagiarism detection and recommender systems. With increasingly larger data sets, scalability has become an important issue. Many of the most space and running time efficient algorithms are based on locality sensitive hashing. The de facto standard approach to quickly answer nearest-neighbor queries on such a data set is usually a form of min-hashing. Not only is min-hashing very fast, but it is also space efficient and can be implemented in many computational models aimed at dealing with large data sets such as MapReduce and streaming. A significant drawback is that minhashing and related methods are only able to handle insertions to user profiles and tend to perform poorly when items may be removed. We initiate the study of scalable locality sensitive hashing (LSH) for dynamic data-streams. Specifically, using the Jaccard index as similarity measure, we design (1) a nearest-neighbor datastructure maintainable in dynamic data streams and (2) a sketching algorithm for similarity estimation. Our algorithms have little overhead in terms of running time compared to previous LSH approaches for the insertion streams, and drastically outperform previous algorithms in case of deletions

EPRO DM - 024	Privacy-Preserving Social Media Data Publishing for Personalized Ranking-Based Recommendation

Personalized recommendation is crucial to help users find pertinent information. It often relies on a large collection of user data, in particular users' online activity (e.g., tagging/rating/checking-in) on social media, to mine user preference. However, releasing such user activity data makes users vulnerable to inference attacks, as private data (e.g., gender) can often be inferred from the users' activity data. In this paper, we proposed PrivRank, a customizable and continuous privacy-preserving social media data publishing framework protecting users against inference attacks while enabling personalized ranking-based recommendations. Its key idea is to continuously obfuscate user activity data such that the privacy leakage of user-specified private data is minimized under a given data distortion budget, which bounds the ranking loss incurred from the data obfuscation process in order to preserve the utility of the data for enabling recommendations. An empirical evaluation on both synthetic and real-world datasets shows that our framework can efficiently provide effective and continuous protection of user-specified private data, while still preserving the utility of the obfuscated data for personalized ranking-based recommendation. Compared to state-of-the-art approaches, PrivRank achieves both a better privacy protection and a higher utility in all the ranking-based recommendation use cases we tested.

EPRO DM - 025	GERF: a group event recommendation framework based on learning-to-rank

Event recommendation is an essential means to enable people to find attractive upcoming social events, such as party, exhibition and concert. While growing line of research has focused on suggesting events to individuals, making event recommendation for a group of users has not been well studied. In this paper, we aim to recommend upcoming events for a group of users. We formalize group recommendation as a ranking problem and propose a group event recommendation framework GERF based on learning-to-rank technique. Specifically, we first analyze different contextual influences on user's event attendance, and extract preference of user to event considering each contextual influence. Then, the preference scores of the users in a group are taken as the features for learning-to-rank to model the preference of the group. Moreover, a fast pairwise learning-to-rank algorithm, Bayesian group ranking, is proposed to learn ranking model for each group. Our framework is easily to incorporate additional contextual influences, and can be applied to other group recommendation scenarios. Extensive experiments have been conducted to evaluate the performance of GERF on two real-world datasets and demonstrate the appealing performance of our method on both accuracy and time efficiency.

EPRO DM - 026	Tour Sense: A Framework for Tourist Identification and Analytics Using Transport Data

We advocate for and present TourSense, a framework for tourist identification and preference analytics using city-scale transport data (bus, subway, etc.). Our work is motivated by the observed limitations of utilizing traditional data sources (e.g., social media data and survey data) that commonly suffer from the limited coverage of tourist population and unpredictable information delay. TourSense demonstrates how the transport data can overcome these limitations and provide better insights for different stakeholders, typically including tour agencies, transport operators and tourists themselves. Specifically, we first propose a graph-based iterative propagation learning algorithm to recognize tourists from public commuters. Taking advantage of the trace data from the identified tourists, we then design a tourist preference analytics model to learn and predict their next tour, where an interactive user interface is implemented to ease the information access and gain the insights from the analytics results. Experiments with real-world datasets (from over 5.1 million commuters and their 462 million trips) show the promise and effectiveness of the proposed framework: the Macro and Micro F1 scores of the tourist identification system achieve 0.8549 and 0.7154 respectively, whereas the tourist preference analytics system improves the baselines by at least 25.53% and 11.44% in terms of precision and recall.

EPRO DM - 027	Model-Based Synthetic Sampling for Imbalanced

Imbalanced data is characterized by the severe difference in observation frequency between classes and has received a lot of attention in data mining research. The prediction performances usually deteriorate as classifiers learn from imbalanced data, as most classifiers assume the class distribution is balanced or the costs for different types of classification errors are equal. Although several methods have been devised to deal with imbalance problems, it is still difficult to generalize those methods to achieve stable improvement in most cases. In this study, we propose a novel framework called model-based synthetic sampling (MBS) to cope with imbalance problems, in which we integrate modeling and sampling techniques to generate synthetic data. The key idea behind the proposed method is to use regression models to capture the relationship between features and to consider data diversity in the process of data generation. We conduct experiments on thirteen datasets and compare the proposed method with ten methods. The experimental results indicate that the proposed method is not only comparative but also stable. We also provide detailed investigations and visualizations of the proposed method to empirically demonstrate why it could generate good data samples.

EPRO DM - 028	Predicting Consumption Patterns with Repeated and Novel Events

There are numerous contexts where individuals typically consume a few items from a large selection of possible items. Examples include purchasing products, listening to music, visiting locations in physical or virtual environments, and so on. There has been significant prior work in such contexts on developing predictive modeling techniques for recommending new items to individuals, often using techniques such as matrix factorization. There are many situations, however, where making predictions for both previously-consumed and new items for an individual is important, rather than just recommending new items. We investigate this problem and find that widely-used matrix factorization methods are limited in their ability to capture important details in historical behavior, resulting in relatively low predictive accuracy for these types of problems. As an alternative we propose an interpretable and scalable mixture model framework that balances individual preferences in terms of exploration and exploitation. We evaluate our model in terms of accuracy in user consumption predictions using several real-world datasets, including location data, social media data, and music listening data. Experimental results show that the mixture model approach is systematically more accurate and more efficient for these problems compared to a variety of state-of-the-art matrix factorization methods.

EPRO DM - 029	Multi-view Scaling Support Vector Machines for Classification and Feature Selection

With explosive growth of data, the multi-view data is widely used in many fields, such as data mining, machine learning, and computer vision and so on. Because such data always has complex structure, i.e. many categories, many perspectives of description and high dimension, how to formulate an accurate and reliable framework for multi-view classification is a very challenging task. In this paper, we propose a novel multi-view classification method by using multiple multi-class support vector machines (SVMs) and a novel collaborative strategy. Here each multi-class SVM integrates the scaling factor to renewedly adjust the weight allocation which is beneficial to highlight some more discriminative features. Furthermore, we use the decision function values of multiple learners to combine multiple multi-class learners, and then determine the final classification results according to a final confidence score. In addition, through a series of theoretical analyses, we bridge the proposed model with the solvable problem and solve it by an iterative optimization method with convergence. We evaluate the proposed method on several image datasets and face datasets, and the experimental results demonstrate that our proposed method performs better than other state-of-the-art learning algorithms.

EPRO DM - 030	Multi-objective Transport System Based on Regression Analysis and Genetic Algorithm using Transport Data

Intelligent Transportation Systems (ITS) is a cutting-edge traffic solution employing state-ofthe-art information and communication technologies. Optimized bus-scheduling; being an integral part of ITS ensures safety, efficiency, traffic congestion-reduction, passengers' forecast, resource-allocation, and drivers' experience enhancement. Nevertheless, of its significance, recent years have witnessed limited research carried out in this context. In this paper, we apply a uni-variate multi-linear regression over the past three years of data from a renowned bus company and forecasted potential passengers for different days in a week. Moreover, a minimum number of different type of buses have been calculated and bus optimization has been performed in a Genetic Algorithm. The results accurateness has been validated by using absolute deviation (MAD) and mean absolute percentage error (MAPE). The values of MAD (99.14) and MAPE (8.7.)

EPRO DM - 031	Modeling the Parameter Interactions in Ranking SVM with Low-Rank Approximation

Ranking SVM, which formalizes the problem of learning a ranking model as that of learning a binary SVM on preference pairs of documents, is a state-of-the-art ranking model in information retrieval. The dual form solution of a linear Ranking SVM model can be written as a linear combination of the preference pairs, i.e., $w = \Sigma_{(i,j)} \alpha_{ij} x_i - x_j$), where α_{ij} denotes the Lagrange parameters associated with each preference pair (i,j). It is observed that there exist obvious interactions among the document pairs because two preference pairs could share a same document as their items, e.g., preference pairs (d_1 ,d_2) and (d_1 ,d_3) share the document d_1 . Thus it is natural to ask if there also exist interactions over the model parameters α_{ij} , which may be leveraged to construct better ranking models. This paper aims to answer the question. We empirically found that there exists a low-rank structure over the rearranged Ranking SVM model parameters α_{ij} , which indicates that the interactions do exist. Based on the discovery, we made modifications on the original Ranking SVM model by explicitly applying low-rank constraints to the Lagrange parameters, achieving two novel algorithms called Factorized Ranking SVM and Regularized Ranking SVM, respectively. Specifically, in Factorized Ranking SVM each parameter α_{ij} is decomposed as a product of two low-dimensional vectors, i.e., $\alpha_{ij} = \langle v_i, v_j \rangle$, where vectors v_i and v_j correspond to document i and j, respectively; In Regularized Ranking SVM, a nuclear norm is applied to the rearranged parameters matrix for controlling its rank.

EPRO DM - 032	Multi-Party High-Dimensional Data Publishing under Differential Privacy

In this paper, we study the problem of publishing high-dimensional data in a distributed multi-party environment under differential privacy. In particular, with the assistance of a semi-trusted curator, the parties collectively generate a synthetic integrated dataset while satisfying ε -differential privacy. To solve this problem, we present a differentially private sequential update of Bayesian network (DP-SUBN) approach. In DP-SUBN, the parties and the curator collaboratively identify the Bayesian network N that best fits the integrated dataset in a sequential manner, from which a synthetic dataset can then be generated. The fundamental advantage of adopting the sequential update manner is that the parties can treat the intermediate results provided by previous parties as their prior knowledge to direct how to learn N . The core of DP-SUBN is the construction of the search frontier, which can be seen as a priori knowledge to guide the parties to update N . Leveraging the correlations of attribute pairs, we propose exact and heuristic methods to construct the search frontier. In particular, to privately quantify the correlations of attribute pairs without introducing too much noise, we first put forward a non-overlapping covering design (NOCD) method, and then devise a dynamic programming method for determining the optimal parameters used in NOCD. Through privacy analysis, we show that DP-SUBN satisfies ε -differential privacy. Extensive experiments on real datasets demonstrate that DP-SUBN offers desirable data utility with low communication cost.

THANK YOU!