Discovery of College Students in Financial Hardship

Conference Paper · November 2015

DOI: 10.1109/ICDM.2015.49

Chu Guan
University of Science and Technology of China

Xinjiang Lu
Baidu Online Network Technology

Enhong Chen
University of Science and Technology of China

Wenjun Zhou
University of Tennessee

Discovery of College Students in Financial Hardship

Chu Guan*, Xinjiang Lu†, Xiaolin Li‡, Enhong Chen*, Wenjun Zhou¶, Hui Xiong§
‡Nanjing University, China, lixl@nju.edu.cn, *corresponding author*
*University of Science and Technology of China, guanchu@mail.ustc.edu.cn, cheneh@ustc.edu.cn
†Northwestern Polytechnical University, China, xjlu@mail.nwpu.edu.cn
§Rutgers University, USA, hxiong@rutgers.edu
¶University of Tennessee Knoxville, USA, wzhou4@utk.edu

***Abstract*—College students with financial difficulties refer to those whose families can hardly afford their high tuition in universities, and should be supported by modern funding system. Indeed, students' economic plight negatively impact their mental health, academic performance, as well as their personal and social life. While funding students in financial hardship is widely accepted, there is limited understanding and research on effectively identification of the qualifying students. Traditional approaches relying on advisers' personal assessments are inefficient, and such subjective judgements may not reflect the truth. To this end, in this paper, we explore the data mining techniques for identifying students who are qualified for financial support. Specifically, we investigate students' complex behaviors on campus from multiple perspectives, and develop a learning framework, named *Dis-HARD*, by jointly incorporating the heterogeneous features to predict the portfolio of stipends a given student should be awarded. Our framework formalizes the above problem as a multi-label learning problem. Along this line, we first extract discriminative features from three perspectives: (i) smartcard usage behavior, (ii) internet usage behavior and (iii) trajectory on campus. Then, we develop a linear loss function with regularization to solve this multi-label classification problem. In addition, to effectively exploit the students' similarity and label dependency, we incorporate the graph Laplacian and composite $\ell_{2,1}$-norm into the regularization of our model, and develop a re-weighted algorithm to achieve effective optimization. Finally, experiments on real-world data demonstrate that our method consistently provides better performance compared to the existing state-of-the-art methods.**

Keywords*—*Student behavior analysis, sequential pattern mining, multi-label classification

I. Introduction

Students are the main assets of universities, and students' performance plays an important role in producing the best quality graduates and post-graduates, who will enter the workforce of the country as contributors and leaders, thus responsible for the country's economic and social development. Specially, supporting college students in financial hardship is a critical issue for students' families and university administration [1]. Students in financial hardship are more likely to have anxiety and depression [2] [3], which can negatively impact their academic performance, and possibly lead to dropouts [4]. Consequently, students with financial hardship in universities should be a concern not only to the administrators and educators, but also to corporations in the labor market.

Generally speaking, the current funding systems are mostly voucher systems, providing funds to students depending on their circumstances and allowing them to use those funds at the institutions and programs of their choice. Institutions often adopt the "students proclaiming + advisers assessing" scheme to pick the "right" students. Over time, the components of the aid system have multiplied, the rules and regulations associated with these programs have become more elaborate, and the eligibility criteria and application processes have become more complex. However, the process of evaluating students' qualification is usually manually conducted, taking into consideration of students' family background, daily expenditures, and academic performance. Therefore, it is often time-consuming for staffs coping with such case-study type of assessments, and involves much subjective judgements, which might not reflect the truth.

For modern university, there is a clear trend to augment physical facilities with sensing, computing and communication capabilities. Many colleges and universities have established plenty of advanced information management and operation systems, which make students' life and study more convenient and efficient. The digital traces left by students while interacting within cyber-physical spaces are accumulating at an unprecedented breadth, depth and scale, and we call all those traces the "digital footprints" [5]. The explosive growth in the amount of available data has created an opportunity for us to automatically analyze students' behavior using novel information technologies. Furthermore, students behavior analytics in universities has become important for both researchers and developers, because it can reveal the patterns of college students behaviors at the individual, group, and societal scales, and thus enable a variety of applications.

To this end, in this paper, we aim to develop data mining solutions for identifying students in financial hardship by exploiting the "digital footprints" left by students. Indeed, the digital traces of college students encode different behavior patterns that reflect their unique habits, mental statuses and financial conditions. If properly and effectively analyzed, these underlying patterns could be a source of rich intelligence for identifying students in financial hardship. In the literature, while there are some related work analyzing students' digital footprints, such as students' learning performance assessment [6], understanding students' study process [7], and investigating sentimental/psychological status for social goods [8], the problem of discovering students with financial difficulties is still under explored. To fill this gap, we propose to develop a learning framework (named ***Dis-HARD***) to discover the students in financial hardship with heterogeneous behavioral information. In other words, we aim to predict the portfolio of stipends a given student should be offered.

Since the large amount of multi-sourced behavioral data

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.49

has being accumulated, it is difficult to effectively extract valuable information. For instance, with the growing complexity of information management systems at universities, the trajectories extracted from digital traces involve more and more diverse accessing devices, heterogeneous sensors, and different buildings/zones on campus. Such trajectories must be encoded into sequences of discrete event symbols, which makes it challenging to unravel meaningful and significant temporal-spatial structures. In this paper, we examine factors reflecting college students' financial conditions by exploiting their complex behaviors in a "multi-scale" fashion.

Specifically, we investigate the students' behavior from three perspectives: (1) smartcard usage, (2) internet usage within campus, and (3) students' trajectories on campus. First, most college students in modern universities are required to have their own smartcard, which can be used as their IDs (for accessing facilities, etc.) and digital wallets (for making payment and depositing etc.) on campus. Consequently, with the logs of smartcard usage, we can examine students' various activities on campus. Second, if a student spends too much time on internet services for entertainment, such a behavior can provide an evidence to believe that he/she is not working hard, thus may not be qualified to offer for a stipend. In addition, with the manner (wireless or wired) of accessing internet services, we can infer whether he/she uses mobile devices a lot, which is another evidence to judge one's financial conditions. For the above two kinds of information, we inspect and extract features from raw data. Third, by integrating students' smartcard usage and internet usage records, plus the logs of entering/leaving dormitories, we can recover the students' coarse grained trajectories on campus. For the trajectories, we utilize a high-level approach to disaggregate features. That is, given the temporal-spatial sequences constructed, we employ a "skeletonization" [9] approach to transform the raw sequential data into the semantic sequences, which is much easier to discover meaningful patterns encoding students' implicit behavior patterns, and helpful to identify students in financial hardship.

On the other hand, to effectively learn the transformation from students' complex behavior (feature space) to the portfolio of stipends the college students would be offered (label space), we develop a learning approach by exploiting the correlation among students and the correlation of stipends (label dependency). In particular, first, we notice that students in similar financial difficulties would behave alike to each other, e.g. they are more likely to have dinner in school cafeteria due to the low price, and access internet in a wired way as they do not have mobile PCs. If such correlation is incorporated into our learning process, the performance of prediction will be definitely improved. Along this line, by minimizing the product of students' similarity and portfolio distance (i.e. the difference among different portfolio of stipends), we employ the graph Laplacian term to capture such correlation in our approach. Second, to effectively explore the relationship of students' behavior patterns and their corresponding portfolio of stipends, we additionally examine the label dependency (i.e. dependencies among different types of stipends). Intuitively, students exhibiting similar behavior patterns are expected to be offered with similar types of stipends. Besides, due to quota limitation, a student may often obtain a few types of stipends compared to the available types (e.g. 2 types of stipends compared to more than fifty available types of stipends), which gives rise to the issue of sparsity. However, for specific students, the stipends offered are often appeared together, in other words, the portfolios of stipends are not highly diversified. Therefore, we adopt a sparsity regularization term in our model for dimensionality reduction by capturing the relationship among different stipends. More specifically, as the ℓ_1-norm regularization is performed to give sparse solution and ℓ_2-norm based loss function is sensitive to outliers, we utilize an $\ell_{2,1}$-norm regularization [10] in our work. Furthermore, due to the difficulty of solving a composite $\ell_{2,1}$-norm compared to a lasso term, we propose an efficient re-weighted algorithm to solve the resultant convex optimization problem.

Finally, we evaluate the proposed learning framework over a real-world dataset involving more than 13K college undergraduate students, and the experimental results show that our method consistently provides better performance compared to the other state-of-the-art methods.

II. Problem Statement and Framework Overview

In this section, we first present the problem formulation, then introduce the preliminaries followed by framework overview.

A. Problem Formulation

The activities of the college students' daily life in campus are continuously recorded/logged, when they make payment with smartcard, consuming internet services in campus, using their IDs to access some public facilities, or entering/leaving their dormitories. Therefore, discovering college students in financial hardship requires to disaggregate those records left by students into different feature sets (detailed in Section III), and predict the *funds portfolio* (i.e., what types of stipends the students should be offered). Note that, due to the diversity of college students, in this study, our problem focuses on undergraduates.

Formally, given a set of college students $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ and K available types of stipends, the goal of our problem is to predict students' *funds portfolio* according to their levels of financial hardship, i.e. $\mathbf{Y} = (\mathbf{y}_n)_{n=1}^{N} = f(\mathbf{X})$, where $\mathbf{y}_n \in \mathbf{Y}(\mathbf{y}_n \in \mathbb{R}^K)$ denotes the *funds portfolio* student u_n should be offered, and $\mathbf{x}_n \in \mathbf{X}$ ($\mathbf{X} = (\mathbf{x}_n)_{n=1}^{N}$, $\mathbf{x}_n \in \mathbb{R}^M$) is the feature vector extracted from student u_n's behavior. In this study, we assume that each student's financial hardship (or *funds portfolio*) is associated with multiple facets of his/her daily life, including smartcard usage behavior, internet usage behavior and trajectories within campus. Essentially, there are two major tasks: (1) unraveling meaningful features with the "digital footprints" and (2) inferring the *funds portfolio* of students.

B. Preliminaries

Before presenting the overview of our framework, we need to provide some preliminaries in advance.

It is common that the modern universities have more than one campuses, thus there might be various places in different categories (i.e. the functionality of a place) in one university. Formally, the set of place categories in university is denoted by $\mathcal{C} = (c_\zeta)_{\zeta=1}^{Z}$, where Z is the number of place functions (categories) in university. Consequently, we define the places in university as $\mathbb{P} = (p_\lambda)_{\lambda=1}^{\Lambda}$, and each place $p \in \mathbb{P}$ is defined as a tuple $(p.loc, p.fun)$, where $p.loc$ is p's physical location and $p.fun \in \mathcal{C}$ is p's functionality and we use $c(\cdot)$ to denote the place function, i.e. $c(p) = p.fun$. Besides, due to the complex behavioral information of students, we have the set of behavior categories $\mathcal{B} = (b_j)_{j=1}^{J}$. For the sake of clarity, in the rest of this paper, we use u instead of $u_n \in \mathcal{U}$ to generally represent a college student if not particularly indicated.

To effectively inspect the digital traces (including multiple types of records) left by students, we need formal definition for the digital records. Intuitively, a record corresponds to student u's *activity a*, which is formally defined as follows:

Definition 1. Activity. *An activity* a *of student* u *contains a time range $T_a = \{t^a_{start}, t^a_{end}\}$, a place $p \in \mathbb{P}$, a behavior category $b_j \in \mathcal{B}$ and corresponding attributes $a.attr$.*

Note that $a.attr$ is a vector, and the size of $a.attr$ depends on the behavior category $b_j \in \mathcal{B}$ of *activity a*. For example, the corresponding attributes of smartcard usage activity probably differ from internet service usage activity.

Along this line, when examining the historical behavioral records of u, we obtain the sequence of activities $seq(u) = (a_l)_{l=1}^{L}$. Note that, for $a_i, a_j \in seq(u)$, their behavior categories might be different, e.g. a_i could be smartcard usage activity and a_j could be internet service usage activity. Therefore, if we disaggregate $seq(u)$ into different sequences (named *categorized sequence*) according to the behavior category, then student u can have multiple sub-sequences: $seq(u) = \{\tilde{seq}^{(b_1)}(u), \dots, \tilde{seq}^{(b_J)}(u)\}$. Then, $\forall u \in \mathcal{U}$, we extract *categorized sequences* and obtain the following set: $Seq = (\{seq(u_n)\})_{n=1}^{N} = (\{\tilde{seq}^{(b_1)}(u_n), \dots, \tilde{seq}^{(b_J)}(u_n)\})_{n=1}^{N} = \{\tilde{Seq}^{(b_1)}, \dots, \tilde{Seq}^{(b_J)}\}$. We present the above symbols and their brief descriptions in Table II.

Moreover, by examining $\tilde{seq}^{(b_j)}(u) \in seq(u)$, we observe that some activities (in the same category) occur adjacently in a relatively very short time range, and form *session*. Particularly, an *activity* which does not have other nearby neighbors can also be treated as a special session. The formal definition of *session* is as follows:

Definition 2. Session. *A session* s *of student* u *contains a time range $T_s = \{t^s_{start}, t^s_{end}\}$, and γ adjacent activities $\{a_1, \dots, a_\gamma\}$, which satisfies $t^s_{start} = t^{a_1}_{start}$, $t^s_{end} = t^{a_\gamma}_{end}$ and there is no other session s^* that makes $T_s \subseteq T_{s^*}$. Meanwhile, $\forall i \in [1, \gamma)$, we have $(t^{i+1}_{start} - t^i_{end}) < \delta$, $dis(p_{i+1}, p_i) < \phi$, where δ and ϕ are two predefined time and distance thresholds for merging activities, and $dis(p_{i+1}, p_i)$ indicates the distance between p_{i+1} and p_i.*

Intuitively, the *session* of a student represents his/her period of engaging at least one consecutive activities with same behavior category, e.g. while a student having lunch, he/she probably makes multiple payments for meal, drinks, and dessert etc., thus he/she can have adjacent smartcard usage activities within, say, one hour.

C. Framework Overview

The overview of our proposed prediction (multi label classification) framework (***Dis-HARD***) is illustrated in Fig.1, which mainly consists of two modules (or steps):

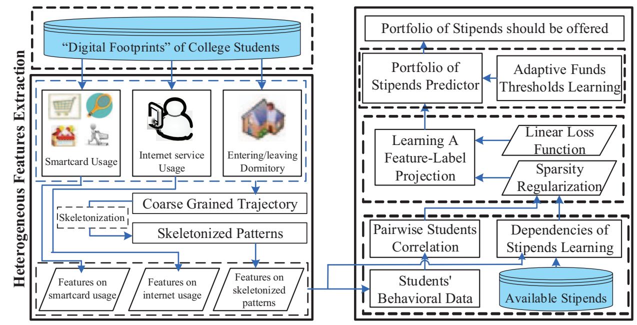


Fig. 1. The framework overview of ***Dis-HARD***.

Step 1: Heterogeneous Features Extraction and Integration. Given $\mathcal{U}$, we first collect the students' various digital records/logs, which mainly contain smartcard usage records, internet service usage records and entering/leaving dormitory records. Then, we prune the meaningful fields of each kind of records, and assembly the set of *categorized sequences* (Seq) for these three types of behavior records. After the set of *categorized sequences* being obtained (i.e., $\{\tilde{Seq}^{(Smartcard)}, \tilde{Seq}^{(Internet)}, \tilde{Seq}^{(Dormitory)}\}$ denoted by Seq'), we first construct and distill semantic trajectories $Tra^{(Semantic)}$ from Seq' by employing "skeletonization" approach, then extract three feature sets $\mathbf{X}^{Smartcard}$, $\mathbf{X}^{Internet}$ and $\mathbf{X}^{Tra}$ by exploiting $\tilde{Seq}^{(Smartcard)}$, $\tilde{Seq}^{(Internet)}$, and $Tra^{(Semantic)}$ respectively.

Step 2: *Funds Fortfolio* Predicting. After features being extracted from different perspectives, we feed all these features to a linear regression model with a regularization term, which incorporates both the students' behavior patterns correlation and the dependency among different funds. The $\ell_{2,1}$-norm regularization term utilized in our approach can penalize the regression coefficients with respect to the significance of different features, and cope with sparsity while incorporating the label dependency. Based on this, we can learn a *funds portfolio* predictor (thus can estimate the financial hardship) by minimizing the loss function. Furthermore, the predicted *funds portfolio* should be a binary vector, therefore, we develop an adaptive thresholds learning method for determining such *portfolios*. Finally, we infer the *funds portfolio* for new coming student with learned feature-label projection.

III. Feature Extraction

Due to the complexity of information management systems deployed in universities, and the diversity of college students' behavior, we inspect the students' behavior in a "multi-scale" fashion. In the following subsections, we describe the details of each feature set.

A. Smartcard Usage Behavior

Towards understanding the situations of college students' financial hardship, given student u, we examine and calculate the following attributes of *activities* in $\tilde{seq}^{(Smartcard)}(u)$: account balance (*Bal*), accumulative number of records of using smartcard (*ANR*), the number of records within one day (*Daily*) and the amount of money involved in each transaction (*TransVol*). The above four attributes are examined from *activity* (Def. II-B) perspective, we additionally inspect attributes from *session* perspective, i.e. we examine the number of *sessions* within one day (*Daily.s*) and the sum of *TransVol* in one session (*TransVol.s*). We subsequently obtain the following sequences of mentioned attributes: (a) $ts_{(Bal)}$, (b) $ts_{(ANR)}$, (c) $ts_{(Daily)}$, (d) $ts_{(TransVol)}$, (e) $ts_{(Daily.s)}$ and (f) $ts_{(TransVol.s)}$, which constitute the set of sequences disaggregated from smartcard usage behavior, i.e. $TS^{(Smartcard)}$. The feature set $\boldsymbol{X}^{Smartcard}$ is introduced as follows:

1) Overall Descriptive Features: Generally, the sequences in $TS^{(Smartcard)}$ exhibit the temporal dynamics in student's smartcard usage behavior, thus can depict the student's financial situations to some extent. It is vital to exploit overall descriptive statistics, as they can describe the basic properties of distribution from multiple aspects. Given a sequence in $TS^{(Smartcard)}$, we extract the first order and second order descriptive statistics (i.e., minimum, maximum, median, mean, 1st-quartile, 3rd-quartile, standard deviation, skewness, and kurtosis) as features.

2) Self-defined Features: Besides, we define some features with meaningful semantics.

Length of longest decreasing subsequences: We examine the longest decreasing subsequences of $ts_{(Bal)}$, and use the length of this subsequence as a feature to describe the student's financial endurance.

Hopping count: We count the elements in $ts_{(Bal)}$ with values greater than their next elements, and take this number to capture the frequency of deposit activities.

Length of longest silent period: For the sequence $ts_{(Daily)}$, we take the longest subsequence with zero value to describe the number of days without using smartcard.

Deposit amount: Sum up the amount of deposits by exploiting $ts_{(TransVol)}$, and the way of identifying deposit activity is same as extracting *Hopping count*.

Ratio of weekend to weekday: By examining $ts_{(TransVol)}$, we calculate the ratio of transaction amount on weekends to weekdays.

Features from $ts_{(ANR)}$: With $ts_{(ANR)}$ reflecting students' historical frequencies of using smartcard, we extract the number of records in the beginning (e.g. the start of one month), and the increased ratio of smartcard usage activities within the time span, like, one month.

B. Internet Service Usage Behavior

The feature set $\mathbf{X}^{Internet}$ is extracted from the following attributes within certain time period (and thus yields sequences): (a) the sequence of internet service usage fees $ts_{(Fee)}$, (b) the sequence of frequencies on daily internet usage $ts_{(Daily_{Int})}$, (c) the sequence of time spent on using internet service $ts_{(Dur)}$ and (d) the sequence of download and upload internet traffics $ts_{(Traffic)}$. Besides, we examine the above four sequences from *session* perspective, which generates $ts_{(Fee.s)}$, $ts_{(Daily_{Int}.s)}$, $ts_{(Dur.s)}$ and $ts_{(Traffic.s)}$. Therefore, we have $TS^{Internet} = \{ts_{(Fee)}, ts_{(Daily_{Int})}, ts_{(Dur)}, ts_{(Traffic)}, ts_{(Fee.s)}, ts_{(Dur.s)}, ts_{(Daily_{Int}.s)}, ts_{(Traffic.s)}\}$, and obtain the overall statistical features $\boldsymbol{X}^{Internet}_{Overall} \subset \boldsymbol{X}^{Internet}$ in a similar way adopted before (in Section III-A). We additionally introduce the "domain specific" features:

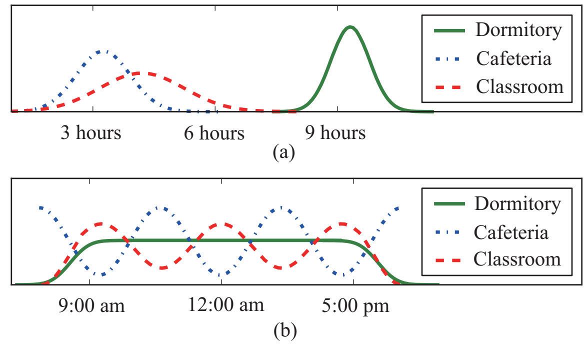


Fig. 2. The distributions of students' population at different places with respect to (a) the time duration, and (b) the time slots in daytime. The y-axis in both figures represents the fraction of students' population.

Longest Silence and the Longest Silence over the upper fee: Based on our observation, the strategies on charging students' using internet services provided by school vary among different institutions, therefore we extract some specific features to reveal students' internet usage patterns when they encountering non-free (but still much cheaper) internet services provided by school. For the university we collected data from, the policy of charging internet service is that, there are two upper bounds, the first is "free" bound under which the internet service is free, and the second is "bonus" bound over which no more internet service usage fees will be charged. Therefore, there is at least one subsequence in $ts_{(Fee)}$ with zero values (named *silence*) in certain time period, and we take the length of longest *silence* as one feature, and adopt the length of longest *silence* exceeding the "bonus" bound as another feature.

Distribution of Time spent on Internet Usage: We examine $ts_{(Dur)}$, and count the time spent on using internet services within campus in different time slots (i.e. morning, afternoon, evening, night, weekday and weekend), and with different accessing manners (wireless or wired), then we obtain the features to describe the students' internet usage temporal distributions.

Distribution of Internet Traffic: Given a student u, the download and/or upload traffic of internet service usage can reflect u's consuming what type of information via internet, and thus judge u's whether or not working hard to some extent. We extract the distributions of u's internet usage traffic with respect to different time periods and ways of accessing internet. Besides, we extract the ratio of upload traffic to download traffic as another feature.

C. Temporal-spatial Trajectory Within Campus

Semantics of different places vary. For instance, if we examine the temporal characteristics of places in campus, students often spend more time staying in dormitory compared to classroom and school cafeteria, meanwhile, more obvious periodic patterns can be observed at classroom and school cafeteria compared to dormitory (Fig.2).

To effectively analyze students' trajectories, we first introduce *semantic trajectory* which is defined as follows:

Definition 3. Semantic Trajectory. *Given a student u, the sequence of activities $seq(u)$, and the place categories $\mathcal{C}$ in university, the semantic trajectory is a sequence of tuples $tra^{(s)}(u) = \langle(p_1, t_1, c(p_1)), \ldots, (p_L, t_L, c(p_L))\rangle$, where $t_i < t_j (i < j)$, $p_i \in \mathbb{P}$ is the i-th place in u's trajectory, and $p_i, p_j \in \mathbb{P}$ can be the same place.*

Note that, in Definition 3, $tra^{(s)}(u)$ is transformed from $seq(u)$, and with this *skeletonization* we can greatly simplify the representation of original trajectory [9].

Due to the periodicity of students' behavior in campus, we utilize a predefined threshold $\triangle t$ to segment $tra^{(s)}(u)$, which yields the set of *skeletonized patterns* : $(\{\vec{sp}_h(u)\})_{h=1}^{H} = (\{c(p_1) \xrightarrow{\triangle t} c(p_2) \xrightarrow{\triangle t} \ldots \xrightarrow{\triangle t} c(p_{\eta_h})\})_{h=1}^{H}$, where $\triangle t$ is the maximum transition time between two adjacent places. Therefore, we have the following definition:

Definition 4. Skeletonized Pattern. *Given $tra^{(s)}(u)$ and $\triangle t$, a* skeletonized pattern *is formulated as $\vec{sp}(u) = c(p_1) \xrightarrow{\triangle t} c(p_2) \xrightarrow{\triangle t} \ldots \xrightarrow{\triangle t} c(p_\eta)$. Meanwhile, the set of* skeletonized patterns *of u is denoted as $SP(u) = (\{\vec{sp}_h(u)\})_{h=1}^{H}$.*

Indeed, there are two types of constraints for a skeletonized pattern: (1) $\triangle t$ ensures temporal continuity, and (2) the semantic constraint ensures semantic consistency.

A major advantage of skeletonized pattern is alleviating the problem of "curse of cardinality" in sequential data analysis. The sequential pattern mining problem has shown to be NP-hard for discrete data [11]. To deal with this problem, some related works reduce the cardinality in pattern mining via a group of operation on the original items [12], [13]. While these methods have been applied in some real-world applications successfully, some emerging issues are addressed when we face the heterogeneous nature of the sequential data. For example, there are different activities in campus life, which generate heterogenous features. In this paper, by identifying meaningful place categories, our method transform the original sequence of places to the sequence of semantic categories, (e.g. school cafeteria and/or dormitory). This is helpful to discover sequential patterns with a higher level of granularity and discover significant knowledge from students' daily life in campus. Our goal is to find frequent skeletonized patterns in campus.

Along this line, if we examine $SP(u)$ for each student $u \in \mathcal{U}$, then we can draw the distribution of all available *skeletonized patterns*, i.e. $P(\mathbb{SP}) = \{\rho_{\vec{sp}_1}, \rho_{\vec{sp}_2}, \ldots, \rho_{\vec{sp}_\Sigma}\}$, where $\mathbb{SP} = \{\vec{sp}_1, \vec{sp}_2, \ldots, \vec{sp}_\Sigma\}$. Moreover, to reveal the significant patterns in students' semantic trajectories, we need the following definition:

Definition 5. Support. *Given a student u_n and the* skeletonized pattern *$\vec{sp}_h(u_n) \in SP(u_n)$, the support of $\vec{sp}_h(u_n)$ is*

$$Sup(\vec{sp}_h) = \begin{cases} \rho_{\vec{sp}_h}, & \textit{if } \vec{sp}_h \in \mathbb{SP} \\ 0, & \textit{otherwise} \end{cases}.$$

TABLE I. TOP-10 FREQUENT SKELETONIZED PATTERNS.

Rank	Top Skeletonized Patterns	Supported by #Trajectories
1	Dormitory	168848
2	Dormitory → Dormitory	69785
3	Dormitory → School Cafeteria	18000
4	Dormitory → Supermarket	15793
5	School Cafeteria	15588
6	Supermarket→Dormitory	15523
7	Dormitory→School Cafeteria→Dormitory	8592
8	Dormitory→Supermarket→Dormitory	7681
9	Dormitory→Teaching Building	5059
10	Dormitory→School Cafeteria→Teaching Building	2417

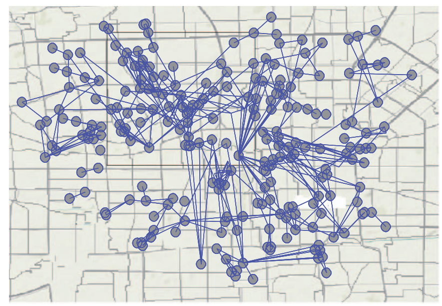


Fig. 3. Semantic Trajectories with a lot of noisy in campus.

To this end, we select top-Q frequent *skeletonized patterns* according to *support* value, i.e. $\mathbb{SP}^{top-Q} = \{\vec{sp}_1^{top-Q}, \vec{sp}_2^{top-Q}, \ldots, \vec{sp}_Q^{top-Q}\}$, and map each *semantic trajectory* to a Q-sized vector, which is denoted as $V_{top-Q}(u) = (\tau(1, SP(u)), \tau(2, SP(u)), \ldots, \tau(Q, SP(u)))$, where $\tau(\cdot)$ is defined as follows ($q \in [1, Q]$):

$$\tau(q, SP(u)) = \begin{cases} Sup(\vec{sp}_q^{top-Q}), & \text{if } \vec{sp}_q^{top-Q} \in SP(u) \\ 0, & \text{otherwise} \end{cases}.$$

Consequently, for all students in $\mathcal{U}$, we extract the Q-sized vector V_{top-Q}, and utilize such vector set as $\mathbf{X}^{(Tra)}$. The main symbols adopted above can be found in Table II.

Finally, we present an example to illustrate the effectiveness of *skeletonization* we employed. Fig.3 and Table I illustrate the original semantic trajectories and encoded *skeletonized patterns* respectively. Here, we set time constraint $\triangle t = 2h$. As can be observed, on the original traces (Fig.3), we can hardly discover meaningful patterns compared to the results illustrated in Table I, which is more insightful.

IV. DISCOVERING STUDENTS IN FINANCIAL HARDSHIP

After feature set $\mathbf{X}$ being extracted, in this section, we describe our approach on how to predict the *funds portfolios* for students by learning a feature-label projection. Our approach is motivated by the following two observations: 1) students, who share similar behavior patterns in campus, tend to have approximative portfolios of stipends, and 2) students in similar financial conditions would behave alike to each other in campus. Therefore, our approach aims to learn the transformation from feature space to label space by exploiting the correlations among students' behavior patterns and the label dependency.

Given the available features of students' behavior patterns $\mathbf{X}$ and the funds portfolio $\mathbf{Y}$, our goal is to find the projection

TABLE II. THE MAIN TERMINOLOGY IN THIS WORK

Symbol	Definition
$\mathcal{U}$	The set of students $\mathcal{U} = \{u_1, \ldots, u_N\}$.
K	# available types of stipends.
$\mathbf{Y}(\in \mathbb{R}^{K\times N})$	Portfolios of stipends for students $\mathcal{U}$.
$\mathbf{X}(\in \mathbb{R}^{M\times N})$	The feature set for students $\mathcal{U}$.
$\mathbb{P}(\in \mathbb{R}^{\Lambda})$	The set of places in university.
$\mathcal{C}(\in \mathbb{R}^{Z})$	The set of place categories in university.
$c(p)$	The place category (functionality) of $p \in \mathcal{P}$.
$\mathcal{B}(\in \mathbb{R}^{J})$	The set of behavioral categories in university.
$seq(u) = (a_l)_{l=1}^{L}$	The sequence of *activities* generated by student u.
$s\tilde{e}q^{(b_j)}(u) \subseteq seq(u)$	The subsequence of $seq(u)$ with behavioral category b_j.
$Seq = (\{seq(u_n)\})_{n=1}^{N}$	The set of sequences generated by students $\mathcal{U}$.
$tra^{(s)}(u)$	The *semantic trajectory* of student u.
$\vec{sp}(u)$	A *skeletonized pattern* segmented from u's semantic trajectory.
$SP(u)$	The set of *skeletonized patterns* of student u, $SP(u) = (\{\vec{sp}_h(u)\})_{h=1}^{H}$.
$\mathbb{SP}$	The set of all available *skeletonized patterns* in $(SP(u_n))_{n=1}^{N}$.
$\mathbb{SP}^{top-Q}$	Top Q frequent skeletonized patterns generated by $\mathcal{U}$.
$V_{Top-Q}(u)$	A Q-sized vector describes the distribution of $SP(u)$ in terms of top-Q frequent *skeletonized patterns*.
$\mathbf{W} \in \mathbb{R}^{M\times K}$	The feature-label projection matrix.
$\mathbf{S} \in \mathbb{R}^{N\times N}$	Pairwise student similarity matrix.
$\mathbf{D} \in \mathbb{R}^{N\times N}$	Diagonal matrix $D_{ii} = \Sigma_j S_{ij}$.
$\mathcal{L}$	The graph Laplacian $\mathcal{L} = \mathbf{D} - \mathbf{S}$.
$\mathbf{R} \in \mathbb{R}^{K\times K}$	The correlation matrix of stipends $\mathbf{R} = (r_{s,t})^{K\times K}$.

$\mathbf{W}$, such that,

$$\begin{aligned}\min_{\mathbf{W}} L(\mathbf{W}) &= \min_{\mathbf{W}} \sum_{n=1}^{N} \|\mathbf{W}^T\mathbf{x}_n - \mathbf{y}_n\|_2^2 + \Omega(\mathbf{W}) \\ &= \min_{\mathbf{W}} f(\mathbf{W}) + \Omega(\mathbf{W})\end{aligned}, \quad (1)$$

where $\mathbf{W} = (w_{i,j})^{M\times K} \in \mathbb{R}^{M\times K}$ captures the relations between students' behavior patterns and the dependencies among stipends, $f(\cdot)$ is the loss function (e.g., least square loss), and $\Omega(\cdot)$ is the regularization term.

We adopt the linear loss function, which is defined as follows:

$$f(\mathbf{W}) = \|\mathbf{Y} - \mathbf{W^T X}\|_F^2.$$

The linear correlation makes our model easy to interpret and, moreover, it leads computational efficiency, compared to hinge loss or logistic loss.

A. Regularization Term $\Omega(\mathbf{W})$

We denote the regularization term of Eq.1 as $\Omega_{\alpha,\beta}(\mathbf{W})$ which aims to capture the intra-correlation of students' behavior patterns and the dependencies among different labels. Here, α and β are two trade-off parameters.

1) Incorporating students similarity: According to the first observation we mentioned before, we aim to optimize:

$$\min_{\mathbf{y}_i,\mathbf{y}_j} \sum_{i,j} s_{i,j} \cdot \|\mathbf{y}_i - \mathbf{y}_j\|_2^2, \quad (2)$$

where $s_{i,j}$ represents the difference between student u_i and u_j, i.e. the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, which is estimated as follows (cosine similarity):

$$s_{i,j} = \frac{\Sigma_{m=1}^{M} x_{i,m} \times x_{j,m}}{\sqrt{\Sigma_{m=1}^{M}(x_{i,m})^2} \times \sqrt{\Sigma_{m=1}^{M}(x_{j,m})^2}}.$$

Then, we utilize the graph Laplacian, i.e. $\mathcal{L} = \mathbf{D} - \mathbf{S}$, to capture the correlation between $\mathbf{x}_i$ ans $\mathbf{x}_j$, where $\mathbf{D} = (d_{i,j})^{N\times N}$ is a diagonal matrix, and its diagonal elements are given by $d_{i,i} = \Sigma_j s_{i,j}$. Subsequently, we have

$$\begin{aligned}&\min_{\mathbf{y_i},\mathbf{y_j}} \sum_{i,j} s_{i,j} \cdot \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 = \min_{\mathbf{y_i},\mathbf{y_j}} 2 \cdot \sum_{i,j} \mathbf{y_i^T}(\mathbf{D} - \mathbf{S})_{\mathbf{i,j}}\mathbf{y_j} \\ &= \min_{\mathbf{y}} 2 \cdot \mathrm{Tr}(\mathbf{y}(\mathbf{D} - \mathbf{S})\mathbf{y^T}) = \min_{\mathbf{y}} 2 \cdot \mathrm{Tr}(\mathbf{y}\mathcal{L}\mathbf{y}^T) \\ &= \min_{\mathbf{W}} 2 \cdot \mathrm{Tr}(\mathbf{W^T x}\mathcal{L}\mathbf{x^T W})\end{aligned} \quad (3)$$

Therefore, instead of optimizing Eq.2 directly, we can optimize Eq.3 in our model.

2) Incorporating label dependency: To capture the dependencies among different labels (stipends), we introduce the second part of the regularization.

First of all, the correlation matrix of funds is denoted by $\mathbf{R} = (r_{s,t})^{K\times K}$, then $\mathbf{R} = \mathbf{R^T}$, and,

$$r_{s,t} = \frac{\Sigma_i y_{s,i} y_{t,i}}{\sqrt{\Sigma_i y_{s,i}^2}\sqrt{\Sigma_i y_{t,i}^2}}.$$

If $r_{s,t} > \epsilon$ (ϵ is a predefined threshold), then stipends s and t are frequently occurred in-pairs. Then we put those highly correlated stipends in a cohesive group and make them co-occur in the regularization term.

In particular, a groups of stipends can be represented by a subset $g \subset \{1, \ldots, K\}$, and can be deemed as one possible portfolio of stipends. Then the power set of $\{1, \ldots, K\}$, i.e. $\mathcal{P}(\{1, \ldots, K\}$, represents all possible portfolios. For instance, a possible portfolio g could be $\{1, 2, 4\}$, where 1, 2 and 4 represent the no.1, no.2 and no.4 available stipends, which is denoted as follows

$$I_g = [1, 1, 0, 1, 0, \ldots, 0].$$

Besides, we construct *group variable* $\mathbf{W}_g \in \mathbb{R}^{M\times K}$, where only entries in group g are preserved and other entries are set to be zero. For instance, if $g = \{2, 6\}$, then

$$\mathbf{W}_g = (\mathbf{0}, \mathbf{w_2}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{w_6}, \mathbf{0}, \ldots, \mathbf{0_{(K)}}).$$

Moreover, to alleviate the sparsity, we utilize a composite $\ell_{2,1}$-norm on the correlated stipends (labels). Recently, sparsity-induced $\ell_{2,1}$-norm is suggested helpful to capture the most significant features associated with class labels [14]. Therefore, we employ $\ell_{2,1}$-norm on each *group variable* $\mathbf{W}_g$ to select discriminant features in each group for *funds portfolio* prediction. Formally, $\ell_{2,1}$-norm of a matrix $\mathbf{W} \in \mathbb{R}^{M\times K}$ is defined as $\|\mathbf{W}\|_{2,1} = \Sigma_{i=1}^{M}\sqrt{\Sigma_{j=1}^{K} w_{i,j}^2}$, then

$$\|\mathbf{W}_g\|_{2,1} = \Sigma_{i=1}^{M}\sqrt{\Sigma_{j=1}^{K}(w_g)_{i,j}^2}.$$

Consequently, we use the term $\Sigma_g\|\mathbf{W}_g\|_{2,1}$ to capture label dependency.

After combining these two regularization terms above, $\mathrm{Tr}(\mathbf{W^T x}\mathcal{L}\mathbf{x^T W}$ and $\Sigma_g\|W_g\|_{2,1}$, we have $\Omega_{\alpha,\beta}(\mathbf{W})$, i.e.,

$$\Omega_{\alpha,\beta}(\mathbf{W}) = \alpha \cdot \mathrm{Tr}(\mathbf{W^T x}\mathcal{L}\mathbf{x^T W}) + \beta \cdot \sum_g \|W_g\|_{2,1}. \quad (4)$$

B. Optimization

Since all terms in Eq.1 are convex, we can obtain a global optimal solution. In this section, we present an iteratively reweighted approach to solve Eq.1.

The main problem of solving Eq.1 is how to tackle the regularization term $\Omega_{\alpha,\beta}(\mathbf{W})$. Generally, it is much more difficult to solve the composite norm $\ell_{2,1}$ than lasso term.

Instead of optimizing Eq.1 directly, we propose to optimize objective function formulated as:

$$J(\mathbf{W}) = f(\mathbf{W}) + \alpha \cdot Tr(\mathbf{W^T x \mathcal{L} x^T W}) + \beta \cdot Tr(\sum_g \mathbf{W}^T \mathcal{F}_g \mathbf{W}_g), \tag{5}$$

where $\mathcal{F}_g \in \mathbb{R}^{M \times M}$ is a diagonal matrix which encodes the composite group $\ell_{2,1}$ structure information and the diagonal elements are given by:

$$\mathcal{F}_{ii} = \frac{1}{2\|\mathbf{w}_g^i\|_2}, \tag{6}$$

where $\mathbf{w}_g^i$ is the i-th row of $\mathbf{W}_g$ $(1 \le i \le M)$. We aim to find an optimal global solution for $\mathbf{W}$. Notice that the minimization of $\mathbf{W}$ depends on $\mathcal{F}_g$ and the computation of $\mathcal{F}_g$ depends on $\mathbf{W}$. Thus, we propose an iteratively re-weighted algorithm. In each iteration, $\mathbf{W}$ is updated along its gradient descent direction until the algorithm converges.

Here, we denote each column of $\mathbf{W}$ as $\mathbf{w}_k$ $(1 \le k \le K)$. By taking the derivative of Eq.5 with $\mathbf{w}_k$ and setting $\partial J / \partial w = 0$, we have

$$\mathbf{X}\mathbf{X}^T \mathbf{w}_k + \alpha \mathbf{x} \mathcal{L} \mathbf{x}^T \mathbf{w}_k + \beta \sum_g \theta(I_g^k = \mathbf{1}) \mathcal{F}_g \mathbf{w}_k = \mathbf{X}(\mathbf{y}^k)^T, \tag{7}$$

where $\theta(\cdot)$ is an indicator function and $\mathbf{y}^k$ is k-th row of $\mathbf{Y}$. $\theta(I_g^k = 1)$ is used to index the group g such that the k-th element in group g is 1. In this way, we get the optimal solution of each $\mathbf{w}_k$ $(1 \le k \le K)$,

$$w_k = (\mathbf{X}\mathbf{X}^T + \alpha \mathbf{x} \mathcal{L} \mathbf{x}^T + \beta \sum_g \delta(I_g^k = 1) \mathcal{F}_g)^{-1} \mathbf{X}(\mathbf{y}^k)^T. \tag{8}$$

In t-th iteration, we update $\mathbf{w}_k$ via Eq.8 and update $\mathcal{F}_g$ via Eq.6, until the algorithm converges. Then we can obtain the global optimal solution of Eq.5 which is also the global optimal solution of Eq.1.

C. Funds Portfolio *Inference*

After obtaining projection matrix $\mathbf{W}$, we can construct the regressor for discovering the students in financial hardship, i.e. predicting the *funds portfolio* the student u_n should be offered ($\tilde{\mathbf{y}}_n = \mathbf{W}^T \mathbf{x}_n$). For a new coming student u_n, we may predict his extent of financial hardship according to the *funds portfolio* he will obtain. Since the output of regressor is continuous, we need to estimate the binary value as follows:

$$y_{k,n} = \begin{cases} 1, & if \quad \tilde{\mathbf{y}}_{k,n} \geqslant \mu_k \\ 0, & if \quad \tilde{\mathbf{y}}_{k,n} < \mu_k \end{cases},$$

where $\tilde{\mathbf{y}}_{k,n}$ is the predictive value for u_n being offered by k-th stipend, and μ_k is the threshold for k-th stipend. If we calculate $1-norm$ of $\mathbf{y_n}$, theoretically we have $0 \le \|\mathbf{y_n}\|_1 \le K$, while in real world, $\|\mathbf{y_n}\|_1 = 0$ is much common and $\|\mathbf{y_n}\|_1 = K$ is almost impossible. Note that the threshold μ_k vary accordingly with respect to the type of stipend.

TABLE III. STATISTICS OF THE EXPERIMENTAL DATA.

Data Sources	Properties	Statistics
General	# Students	13,819
	# Students offered by stipends	3,791
	# Types of Stipends	37
	# Locations in university	2,233
	# Place Categories	15
	Time Period	09.2013 - 09.2014
Smartcard usage	# Records	1,851,935
	# Features	54
Internet usage	# Records	1,887,262
	# Features	115
Entering/leaving dormitory	# Records	1,588,216
Trajectory constructed	Max length of trajectory	175
	# total entries involved	3,963,604

D. Adaptive Thresholds Of μ_k Based On F_1-score

To improve the predictive performance, we adaptively learn the specific thresholds for different types of stipends. In particular, we adopt the metrics defined in [15] to evaluate the prediction performance. For multi-label classification, besides F_1-score, we also consider the macro-average F_1 score and micro-average F_1 score (detailed in Section V-D). For class k, let TP_k, FP_k, TN_k, FN_k be the true positive, false positive, true negative and false negative respectively. Then, we have $\mathrm{MicroPrecision_k}$: $P_k = \frac{TP_k}{TP_k + FP_k}$, $\mathrm{MicroRecall_k}$: $R_k = \frac{TP_k}{TP_k + FN_k}$, $MicroF1_k$: $F1_k = 2\frac{P_k \times R_k}{P_k + R_k}$.

Formally, with the predicted values $\mathbf{W}^{\mathrm{T}} \mathbf{x}_n$ on training dataset, we can evaluate the precision and recall for each stipend. If μ_k is the threshold value for predicting k-th stipend, then we have MicroPrecision $P_k(\mu_k)$ and MicroRecall $R_k(\mu_k)$. Thus, we aim to find μ_k which maximize $\mathrm{MicroF1_k}$ as follows:

$$\max_{\mu_k} \mathrm{MicroF1_k} = \max_{\mu_k} \frac{2}{\frac{\gamma}{P_k(\mu_k)} + \frac{1-\gamma}{R_k(\mu_k)}},$$

where γ is a trade-off parameter that reflects the weights placed precision and recall. In our work, we set $\gamma = 0.5$.

V. Experimental Results

Here we demonstrate our empirical evaluation of the proposed method based on real-world students' data collected from one university in China.

A. The Experimental Data

Table III shows the detailed statistics of our real world data sets. Generally, our data sets include the records of smartcard usage, internet usage and entering/leaving dormitory. After being processed, the data sets cover 13,819 students (undergraduates). Besides, the records of students been offered by stipends in this university were collected. Each entry includes the profile of each student who obtained the stipends, and some basics about the funds (funds name, level and amount). Note that, each student can be offered by multiple types of stipends.

B. Baseline Algorithms

To show the effectiveness of our method, we compare our method against the following state-of-the-art algorithms:

- **Support Vector Machine (SVM)**: Support Vector Machine is a typical maximum-margin classifier. In

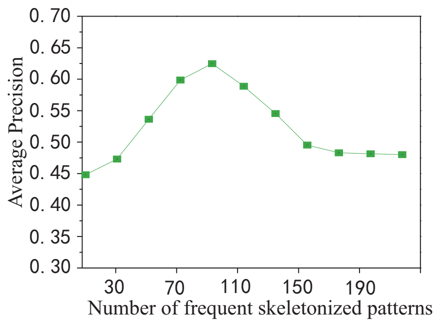


Fig. 4. The performance of *funds portfolio* prediction against the length of V_{top-K} .

order to tackle the multi-label prediction problem, we use one against all strategy to predict the *funds portfolio* with respect to each label. We implemented this algorithm by employing an open sourced software LIBSVM [16].

- **Multiple Kernel Learning (MKL)**: Multiple Kernel Learning assigns a group of base kernels for the problem of feature fusion. One can train a classifier for each type of features and combine the predicted results by utilizing the voting strategy [17].
- **Multi-Label LSI**: Multi-Label Informed Latent Semantic Indexing is a supervised multi-label learning method which preserves the input structure and capture the correlations between multiple outputs simultaneously [18].
- **TODMIS**: TODMIS is a general framework for trajectory community discovery. TODMIS integrates multi-source information into raw trajectory and thus can be used to discover communities [19]. We employ TODMIS to discover students in financial difficulties as another baseline method.

C. Experimental Settings

In Section III-C, we investigate students' trajectories with *skeletonization* and select top frequent *skeletonized patterns* to encode each student's trajectory into a Q-sized vector (V_{Top-Q}). Here, we need to determine the value of Q for our evaluation. We feed $\mathbf{X}^{Tra}$ plus $\mathbf{X}^{Smartcard}$ and $\mathbf{X}^{Internet}$ to our model, and evaluate the predictive performance in terms of Average Precision. As Fig.4 shows that, our model achieves best performance at $Q = 96$, therefore, we set $Q = 96$ in our experiments.

In the regularization of our model, we have two parameters α and β, and we search the optimal settings of α and β within the range of [-50,50] by minimizing the loss function (Eq.1).

Besides, we provide the detailed set up for the baseline methods. (1) SVM[1]: Since SVM cannot directly adopted in multi-label prediction problem, we use one against all strategy to predict *funds portfolio*, i.e. for each type of stipend, we predict whether students can obtain the specific type of stipend, and subsequently combine the predicted labels as the predicted portfolios. Besides, a linear kernel is used in SVM and the penalty parameter $C = 1$. (2) MKL[2]: For MKL algorithm, we use ℓ_1-norm to handle different base kernels and a parameter $C = 1$ to tune the regularization. (3) Multi-Label LSI: We utilize the same parameters as described in [18]. (4) TODMIS: We use average kernel weight $\upsilon = 0.2, i = 1, \ldots, 5$ for five base kernels, and we also applied such weights in ℓ_1-MKL.

Finally, we employ five-fold cross validation for our performance evaluation.

D. Evaluation Metrics

Given the set of feature vectors examined from $\mathcal{U}$ and the ground true *funds portfolios* for $\mathcal{U}$, i.e. $\mathbf{X}$ and $\mathbf{Y}$ respectively, and the predicted *funds portfolios* for students $\mathcal{U}$ denoted by $\tilde{\mathbf{Y}} = (\tilde{\mathbf{y}}_n)_{n=1}^N$, we have the following evaluation metrics for our method.

MacroF1. The macroF1 metric calculates the average of F_1 scores on all labels.

$$\mathrm{MacroF1}(\tilde{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{N}\sum_{n=1}^{N}\frac{2 \times \|\tilde{\mathbf{y}}_n \cap \mathbf{y}_n\|_1}{\|\tilde{\mathbf{y}}_n\|_1 + \|\mathbf{y}_n\|_1}$$

MicroF1. The microF1 is the global calculation of F_1 regardless of classes. The F_1-score can take both the precision and the recall into account, thus can be viewed as a harmonic mean of precision and recall.

$$\mathrm{MicroF1}(\tilde{\mathbf{Y}}, \mathbf{Y}) = \frac{2 \times \sum_{n=1}^{N}\|\tilde{\mathbf{y}}_n \cap \mathbf{y}_n\|_1}{\sum_{n=1}^{N}\|\tilde{\mathbf{y}}_n\|_1 + \sum_{n=1}^{N}\|\mathbf{y}_n\|_1}$$

The larger the MicroF1 score, the better the performance of one classification method will achieve.

Hamming Loss. Hamming loss calculates the fraction of wrong labels to the total number of labels.

$$\mathrm{HammingLoss}(\tilde{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\|\tilde{\mathbf{y}}_n \oplus \mathbf{y}_n\|_1$$

where $\oplus$ stands for the symmetric difference of two sets and $\|\cdot\|_1$ denotes the ℓ_1-norm. The smaller the Hamming loss, the better the performance.

E. Comparison of Different Features

In this subsection, we compare the effectiveness of different behavioral features for *funds portfolio* prediction.

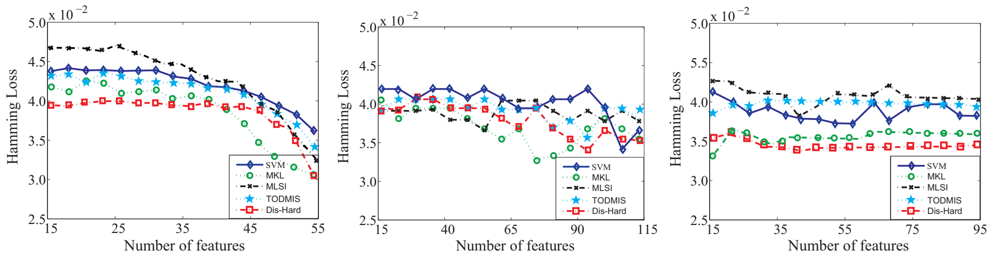
Sensitive to category of features. We divide the features into three subsets according to their behavioral categories (i.e. $\mathbf{X}^{Smartcard}$, $\mathbf{X}^{Internet}$ and $\mathbf{X}^{Tra}$) and investigate which kind of features is more effective. Due to the space limit, we only present the results on *Hamming loss* which is arguably the most important multi-label evaluation criterion. It can be found in Fig.5 that all the extracted features achieve good performances, yet there are features which are substantially better than others. For example, the smardcard usage features perform best with Hamming loss < 0.04 and consistently achieve the best classification results on our method (***Dis-HARD***). The features of internet usage hold the second place of overall in students'

[1] http://www.csie.ntu.edu.tw/cjlin/libsvm/

[2] http://doc.ml.tu-berlin.de/nonsparse_mkl/

TABLE IV. PREDICTIVE PERFORMANCE COMPARISON.

Method	Macro Precision	Macro Recall	Macro Accuracy	Macro F1	Micro Precision	Micro Recall	Micro Accuracy	Micro F1	Hamming Loss
SVM	0.1801	**0.8232**	0.9522	0.3685	0.0125	0.0128	0.9524	0.0193	0.0595
MKL	0.7873	0.5046	**0.9847**	0.6262	0.0495	0.0402	0.9672	0.0438	0.0303
MLSI	0.7363	0.5882	0.9631	0.6583	0.0683	0.0692	0.9642	0.0593	0.0652
TODMIS	0.7692	0.5272	0.9652	0.6272	0.0482	0.0492	0.9662	0.0492	0.0341
Dis-HARD	**0.8319**	0.6331	0.9597	**0.6942**	**0.0826**	**0.0813**	**0.9721**	**0.0872**	**0.0293**



(a) Smartcard usage (b) Internet usage (c) Trajectory

Fig. 5. Average accuracy of different features on the student behavior dataset.

behavior data. In sum, smardcard usage features and internet usage features perform better than trajectory features. Besides, MKL and ***Dis-HARD*** achieve the best performance with all the three types of features. This is because both MKL and ***Dis-HARD*** exploit the intra-correlation of stipends compared to other methods. Moreover, ***Dis-HARD*** achieves the best performance on trajectory features by exploring the semantics of students' behavior. One possible reason is that the students' financial conditions are highly correlated with their smardcard usage behavior and internet usage behavior which can directly reflect their consumption behavior to some extent.

Sensitive to number of features. As shown in Fig.5, we also investigate the performance of the compared methods under different number of features. In Fig.5(a) and Fig.5(b), we can observe that with increasing the number of smartcard usage features, the classification error drops. Besides, as shown in Fig.5(c), the trajectory features are actually homogeneous and generated from the vector $\mathbf{V}^{Top-Q}$, thus the number of trajectory features does not significantly affect the classifier's performance.

Furthermore, by examining the results in Table IV (which evaluate the performance with all features), we can see that the combination of all the features achieves best results in *Hamming Loss*. In sum, the results validate the effectiveness of our feature design and the combination of all features.

F. The Evaluation on Funds Portfolio Prediction

Here, we report our evaluation results of our method comparing to the baseline algorithms. We show both macro and micro measurements in terms of precision, recall, accuracy and F1 score, plus Hamming loss in Table IV.

As observed in table IV, the performance of our method is consistently improved in terms of the F1 measure, accuracy and Hamming loss. The experimental results further confirm our intuitions, i.e. incorporating the students' correlation and the label dependency (dependencies among different types of stipends) into the learning algorithm can indeed improve the predictive performance. On the contrary, SVM considers each student's activity independently, and its performance is generally not as good as others. When we combine its precision and recall together, the F1 measure of our method is consistently improved compared to the baselines. To consider both the global statistical information and individual information, MKL and TODMIS adopt kernels to avoid the curse of dimensionality and measure the similarity among features. These two kernel based learning methods validate the benefits of using multiple sourced information, and thus outperform the performance of the single kernel method (SVM).

VI. RELATED WORK

The main research topic on exploring students' behavioral data is about educational data mining (EDM). Many learning analytics and approaches of EDM has being emerged recently [20], which is applied to examine students' learning process and improve their study performance. To the best of our knowledge, we are the first to study the problem of identifying students in financial difficulties by inspecting students' behavioral data.

Another related work is sequential pattern mining which has been extensively studied in transactional data. Agrawal *et al.* [21] first introduce this problem and adopt Apriori to discover meaningful patterns from sequential data. To handle trajectory data, several pioneer studies [22]–[24] have been employed for sequential pattern mining in spatio-temporal databases. However, due to the temporal and spatial continuity, rigid space partitioning is not suitable for mining useful sequential patterns. Besides, to deal with the sharp boundary problem, Giannotti *et al.* [25] propose the definition of T-pattern, which is a Region-of-Interest (ROI) sequence with temporal annotations, in a collection of GPS trajectories. Moreover, there are some related studies on mining sequential patterns in semantic trajectories. They view each place as an item and define a sequential pattern as a sequence of semantic labels (e.g., home $\rightarrow$ office). For example, in [9], [26], [27], the researchers propose and implement several approaches of temporal skeletonization to cope with the problem of the overwhelming scale and the heterogeneous nature of real-world data.

Although multi-label classification has being attracted a lot of attentions in recent years, very few research efforts have been made on heterogeneous features. In [19], [28], the authors use different kernel functions to handle different feature representations. Gong *et al.* [29] incorporate the structure of features to investigate correlation between different features for mult-task learning. In these studies, $\ell_{2,1}$-norm regularization has shown to be effective in sparse-based feature selection model [30]. Our work focuses on inspecting the correlation among instances and mining the dependencies between labels to help learn the feature-label projection for multi-label classification.

Furthermore, there are also some studies on location based data mining, e.g. Zheng *et al.* [22] investigate the trajectory data by incorporating users' interests based on visited frequencies at various locations. In [27], a similarity among users is discovered based on the sequence property of their movement behavior, which is also capable of mining correlations among locations.

VII. Conclusion

In this paper, we investigated the problem of predicting students' financial hardship, which is equivalent to the problem of predicting students' *funds portfolios* (i.e. the stipend portfolios students should be offered), and formulated it as the multi-label classification problem. Specifically, we first examined the students' multi-sourced behavior information, and extracted effective features from multiple perspectives, including students' smartcard usage, internet usage and trajectories in campus. In particular, for trajectory related features, we utilized a *skeletonization* method to ravel the meaningful patterns encoded in students' trajectories, and prune the top frequent patterns as a feature vector. Then, we developed a linear loss function based learning approach with regularization term incorporating both students' similarity and dependency between different *funds portfolios*. In addition, we developed a re-weighted optimization method to learn the feature-label projection, and adopted an adaptive thresholds learning method for determining the *funds portfolio*. Finally, the experimental results of performance evaluation on real world data validated the effectiveness of our proposed method.

References

[1] S. Baum and S. Schwartz, "Student aid, student behavior, and educational attainment," George Washington University working paper, Tech. Rep., 2013.

[2] V. K. Lim and T. S. Teo, "Sex, money and financial hardship: An empirical study of attitudes towards money among undergraduates in singapore," *journal of Economic Psychology*, vol. 18, no. 4, pp. 369–386, 1997.

[3] B. Andrews and J. M. Wilding, "The relation of depression and anxiety to life-stress and achievement in students," *Br. J. Psychol.*, vol. 95, no. 4, pp. 509–521, 2004.

[4] R. Roberts, J. Golding, T. Towell, and I. Weinreb, "The effects of economic circumstances on british students' mental and physical health," *J. Am. Coll. Health*, vol. 48, no. 3, pp. 103–109, 1999.

[5] D. Zhang, B. Guo, B. Li, and Z. Yu, "Extracting social and community intelligence from digital footprints: an emerging research area," in *Ubiquitous Intelligence and Computing*. Springer, 2010, pp. 4–18.

[6] S. K. Yadav, B. Bharadwaj, and S. Pal, "Data mining applications: A comparative study for predicting student's performance," *arXiv preprint arXiv:1202.4815*, 2012.

[7] P. Veeramuthu and R. Periasamy, "Application of higher education system for predicting student using data mining techniques," 2014.

[8] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[9] C. Liu, K. Zhang, H. Xiong, G. Jiang, and Q. Yang, "Temporal skeletonization on sequential data: patterns, categorization, and visualization," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1336–1345.

[10] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint 2, 1-norms minimization," in *Advances in Neural Information Processing Systems*, 2010, pp. 1813–1821.

[11] G. Yang, "The complexity of mining maximal frequent itemsets and maximal frequent patterns," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ADM, 2004, pp. 344–353.

[12] B. Goethals, "Frequent set mining," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 377–397.

[13] A. Salam and M. S. H. Khayal, "Mining top- k frequent patterns without minimum support threshold," *Knowledge and information systems*, vol. 30, no. 1, pp. 57–86, 2012.

[14] D. Kong and H. Jin, "Towards permission request prediction on mobile apps via structure feature learning," in *Proceedings of SIAM International Conference on Data Mining (SDM15)*. SIAM, 2015.

[15] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[18] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005, pp. 258–265.

[19] S. Liu, S. Wang, K. Jayarajah, A. Misra, and R. Krishnan, "Todmis: Mining communities from trajectories," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 2013, pp. 2109–2118.

[20] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013.

[21] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*. IEEE, 1995, pp. 3–14.

[22] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 791–800.

[23] H. Cao, N. Mamoulis, and D. W. Cheung, "Mining frequent spatio-temporal sequential patterns," in *Data Mining, Fifth IEEE International Conference on*. IEEE, 2005, pp. 8–pp.

[24] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer, "Semitri: a framework for semantic annotation of heterogeneous trajectories," in *Proceedings of the 14th international conference on extending database technology*. ACM, 2011, pp. 259–270.

[25] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 330–339.

[26] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis *et al.*, "Semantic trajectories modeling and analysis," *ACM Computing Surveys (CSUR)*, vol. 45, no. 4, p. 42, 2013.

[27] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Semantic trajectory mining for location prediction," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2011, pp. 34–43.

[28] R. K. Wong, V. Chu, M. Ghanavati, and A. Hamzehei, "Trajectory analysis based on clustering and casual structures," in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[29] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 895–903.

[30] X. Chang, F. Nie, Y. Yang, and H. Huang, "A convex formulation for semi-supervised multi-label feature selection," in *The Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.