Social Security and Social Welfare Data Mining: An Overview

Longbing Cao, *Senior Member, IEEE*

***Abstract*—The importance of social security and social welfare business has been increasingly recognized in more and more countries. It impinges on a large proportion of the population and affects government service policies and people's life quality. Typical welfare countries, such as Australia and Canada, have accumulated a huge amount of social security and social welfare data. Emerging business issues such as fraudulent outlays, and customer service and performance improvements challenge existing policies, as well as techniques and systems including data matching and business intelligence reporting systems. The need for a deep understanding of customers and customer–government interactions through advanced data analytics has been increasingly recognized by the community at large. So far, however, no substantial work on the mining of social security and social welfare data has been reported. For the first time in data mining and machine learning, and to the best of our knowledge, this paper draws a comprehensive overall picture and summarizes the corresponding techniques and illustrations to analyze social security/welfare data, namely, *social security data mining (SSDM)*, based on a thorough review of a large number of related references from the past half century. In particular, we introduce an SSDM framework, including business and research issues, social security/welfare services and data, as well as challenges, goals, and tasks in mining social security/welfare data. A summary of SSDM case studies is also presented with substantial citations that direct readers to more specific techniques and practices about SSDM.**

***Index Terms*—Data mining, government data mining, public sector, public service, social security data mining (SSDM), social security, social welfare, social welfare data mining.**

I. Introduction

MACHINE learning and data mining are increasingly used in business applications [21], and in particular, in public sectors [94]. A distinct public-sector area is social security and social welfare [121] which suffers critical business problems, such as the loss of billions of dollars in annual service delivery because of fraud and incorrect payments [121], [134]. People working in different communities are increasingly interested in "what do social security data show" [93] and recognize the value of data-driven analysis and decisions to enhance public service objectives, payment accuracy, and compliance [101], [121]. Within the marriage of machine learning and data mining with public sectors, an emerging data mining area is the analysis of social security/welfare data.

Mining social security/welfare data is challenging. The challenges arise from business, data, and the mining of the data. Social security data are very complex, involving all the major issues that are discussed in the data quality and engineering field, such as sparseness, dynamics, and distribution. Key aspects contributing to challenges in mining social security data are many, e.g., 1) specific business objectives in social security and government objectives, 2) specific business processes and outcomes, 3) heterogeneous data sources, 4) interactions between customers and government officers, 5) customer behavioral dynamics, and 6) general challenges in handling enterprise data, such as data imbalance, high dimension, and so on.

Studies on social security issues started in the middle of the 20th century [30], [35]. Since then, many researchers have worked on different topics. The majority of research has been conducted from political [44], [55], [58], [60], [73], economic [7], [11], [30], [35], [67], [91], [95], [98], [109], sociological [58], [59], and regional [2], [10], [29], [43], [52], [60], [70], [75], [85], [87], [104] perspectives, compared with a much delayed effort made on the technical aspects [23], [64], [68], [83], [94]. The main issues involve problem analysis, process and policy modeling, business analysis, correlation analysis, infrastructure development, and emerging data-driven analysis. In contrast with the dominant fact and trend of policy and economy oriented studies, very limited research [23], [39], [68], [69], [103], [105]–[108], [110]–[116] can be found in the literature on mining social security and social welfare data.

Social security data mining (SSDM) seeks to discover interesting patterns and exceptions in social security and social welfare data. From the data mining goal perspective, it aims to handle different business objectives, such as debt prevention. From the data mining task perspective, it involves both traditional data mining methods, such as classification, as well as the need to invent advanced techniques, e.g., complex sequence analysis.

Australia is one of the most developed social welfare countries in the world in terms of government policies, infrastructure, the population of benefit recipients, and the advancement of social security techniques and tools [121]. Since 2004, we have been engaged in conducting data mining for the Australian Commonwealth Government[1] through a series of projects. We have developed models, algorithms, and systems to indentify key drivers, factors, patterns, and exceptions, indicating high risk of

Manuscript received January 14, 2011; revised September 3, 2011; accepted November 2, 2011. This paper was recommended by Associate Editor Z. Wang. This work was supported in part by Australian Research Discovery Grants DP1096218, DP0988016, and DP0773412 and Linkage Grants LP100200774, LP0989721, and LP0775041, as well as contract research projects with Centrelink.

The author is with Advanced Analytics Institute, University of Technology Sydney, Sydney, N.S.W. 2207, Australia (e-mail: longbing.cao-1@uts.edu.au).

Digital Object Identifier 10.1109/TSMCC.2011.2177258

[1] www.centrelink.gov.au

customers, customer circumstance changes, declarations, and interactions between customers and government officers.[2] The findings have proven to be very useful for overpayment prevention, recovery, prediction, and deep understanding of customer activities and intervention, which involve the recovery and prevention of overpayments (also called "debt" when referring to the part of payments to which the recipient is not entitled) for the government. These substantial practices, which have been selected as one of the IEEE's International Conference on Data Mining top ten data mining case studies [16], offer an opportunity for us to widely review the relevant work, deeply explore SSDM in conjunction with real-life applications in Australia and present the overview of SSDM in this paper.

In this paper, rather than focusing on a specific SSDM technique, we aim to draw an overall picture of SSDM by sharing our experience, observations, and lessons learned in both reviewing the related work and conducting real-life SSDM tasks. This is the first paper in this field, to the best of our knowledge, that provides a comprehensive literature review of over 100 references and a substantial framework of SSDM. In particular, the main contributions consist of

1) a thorough literature review of social security research in the last half century, and discussion of different categorizations of the related work;
2) a comprehensive framework of SSDM, discussing the main data mining goals, tasks, and principal challenges in mining patterns in social security data;
3) a summary of several case studies, which involve the development of new and effective algorithms and tools to handle social security data. In particular, we highlight the work on mining debt-targeted patterns, such as debt-targeted positive and negative sequences, sequential classifiers using both positive and negative sequences, and combined association rules by engaging multiple sources of data; and
4) the extension of discussions about mining general public-sector data.

This paper is organized as follows. Section II summarizes the related work on social security research in the past 50 years. In Section III, we briefly introduce social security business and data characteristics. Section IV outlines a framework for SSDM, including the main goals, tasks, and challenges in mining social security data. In Section V, we briefly introduce our real-life assignments in conducting SSDM in Australia by illustrating five case studies and discuss the development of actionable knowledge for business needs. Section VI discusses public-sector data mining based on the lessons learned in conducting SSDM in Australia. We conclude this paper in Section VII.

II. Review on Social Security/Welfare Research

A. Comprehensive Picture

Research on social security and welfare issues started in the mid-20th century [30]. Since then, broad-based issues have been added to the investigation and can be categorized into the following main streams.

1) *Political perspective:* One of the main streams of research investigates the problems, issues, factors, and impact of social security and welfare from public policy [44], [73], social policy [55], [60], administration [89], governance [58], resistance [32], and practice viewpoints [99].
2) *Economic perspective:* Another dominant fact and trend is the exploration of issues and the effect of social security models and factors from the standpoint of econometrics, public economics, and political economy [44]. This involves analysis and discussions about economy [95], earnings [14], [49], [57], rating [80], savings [4], [11], [65], [73], growth [109], privatization [71], reform [55], [56], [98], labor supply [54], [67], multientity relationship analysis [30], [58], [59], [93], and optimal arrangements [7], [35], [53], [91], [92].
3) *Sociological perspective:* Some researchers are concerned about the social effect of social security policies on society, such as lifecycle [74], [93], demographic [1], [11], behavior [84], [90], aging [72], retirement [10], [50], [51], [65], [79], [81], fraud [32], [88], fairness and affordability [76], etc.
4) *Regional perspective:* Researchers from different countries introduce the development of social security in their countries, for instance, Canada [52], India [2], Latin America [70], [85], the U.S. [124], Britain [60], Sweden [29], China [75], [104], Italy [87], Germany [96], France [10], and Europe [43].
5) *Technical perspective:* An emerging trend in social security is the study of technical issues, e.g., infrastructure development [64], knowledge management [83], policy and process modeling, data-driven analysis [23], [68], [69], [103], [105]–[108], [110]–[116], and correlation analysis crossing multiple areas [94], [110].

B. Technical Perspective

From the technical perspective, the main issues that have been addressed in the literature focus on several areas, including problem analysis, process and policy modeling, business-oriented analysis, correlation analysis, infrastructure support, and data-driven analysis.

1) *Problem analysis:* From time to time, we find papers discussing or debating the issues of reform [56], crisis [6], [13], issues for policies [55], privatization [71], uncertainty [92], optimization [97], fraud [32], [88], and effect on economy [30], society, capital market [31], human resources [90], [93], etc.
2) *Process and policy modeling:* Different approaches, e.g., empirical analysis, time-series analysis, quantitative comparative analysis, and equilibrium analysis [42], have been used and developed to design, simulate, and evaluate policy, pension, benefit [7], process and their effects, as well as their optimization, choice [65], and performance rating [62] including accuracy [45].

[2] datamining.it.uts.edu.au/ssdm

3) *Business-oriented analysis:* Key business issues, such as earnings and income, rate, benefit claiming, behavior, retirement, risk, saving, etc., are studied from political, economic, sociological, and technical perspectives.
4) *Correlation analysis:* The relationship between social security and other economic systems have been studied; for instance, the relationship with health affairs, Medicare [76], [90], taxation [8], [34], stock and market [31], [34], [41], as well as with political structure [30], economic development [30], labor force [54], [67], [96], and human capital [40].
5) *Infrastructure support:* Discussions have been conducted on building IT systems, supporting the analysis of social security data, simulating and optimizing processes, policy, performance rates, etc.
6) *Data-driven analysis:* Recently, the value of data and data-driven decisions has been increasingly recognized. Various analysis approaches are being developed to investigate "what do social security data show" [93], e.g., to identify drivers, enablers [83], service patterns [69], linkages [81], demographic [1], behavior [84], change, data measures and composites [123], fraud [88], and risk adjustment [27] from nonrandom selection, stochastic forecast [72], sequential analysis [116], time-series [73], equilibrium analysis [66], data mining, knowledge management [89], microestimation [50], and e-government [64] aspects.

Modern computer systems have been widely used in the social security and welfare sectors since the earliest period of the computerization age. Currently, the use of computers for e-government service in the developed countries has reached a very advanced and comprehensive level. The research from an e-government perspective in the social security and social welfare sectors can be categorized into four main streams: IT infrastructure, operational support system, business support system, and decision support system.

1) *IT infrastructure:* The infrastructure supporting business processes, networking, data storage, human–computer interaction, etc.
2) *Operational support systems:* Systems supporting operations, such as network inventory, provisioning services, configuring network components, privacy, security management, etc.
3) *Business support systems:* Systems offering business interactions with customers, for allowance and benefit delivery, service profiling, debt management, review management [121], etc.
4) *Decision support systems:* Systems supporting decision making, including business integrity management, business intelligence systems, real-time and historical data analysis system, risk analysis and management systems, case management system, and decision-making facilities.

C. Related Work of Social Security Data Mining

The public sector has also kept "the frontier spirit alive in the computer science community" [94]. In particular, data-driven decision has recently been increasingly recognized as one of the most powerful tools to improve government service objectives. However, mining social security/welfare data is an open, new area in the data mining community. To the best of our knowledge, only two groups [3] have involved SSDM, and a very limited number of relevant publications can be found in the literature. In the following, we discuss the UNC group's work and address the practices by the UTS group in Section V-A.

In [38], [39], [68], and [69], a case study was conducted on monthly service data and service variations to detect common patterns of welfare services given over time. The study's authors used a simple sequence analysis method on monthly service administrative databases, which indicates what services were given when, to whom, and for how long. While "common" service procedures can be identified by simply applying a frequent sequence analysis method, it appears that no additional advancement has been made in tackling critical challenges in the data, e.g., mixed transactional data, imbalanced items, and labels. The method only identifies general frequent procedures that are commonsense to business people. No informative and implicit patterns can be identified in this case study. From a business perspective, the identified frequent patterns are not very helpful, since they reflect the actual service arrangements implemented as per policies. Business people want to discover something they do not already know about their business and to develop a deep understanding of why, and how, specific problems face the organization.

In our substantial literature review of SSDM, no additional references have been identified that provide substantial insights for mining social security/welfare data. For this reason, this paper presents a comprehensive overview of SSDM, starting from discussing the characteristics of business and data in Section III, followed by an SSDM framework in Section IV.

D. Retrospection on Mining Social Security Data

Our substantial literature review work and practices in Australia (see Section V) reveal the following observations about the existing research on mining social security/welfare data.

1) It is a very open area in terms of applying and conducting pattern/anomaly discovery on social security data (i.e., SSDM). In the very limited work available from the literature, no such systematic work has been done in terms of drawing an overall picture of SSDM from either the business or technical side, nor in addressing the challenges and opportunities in SSDM.
2) According to our experience in conducting SSDM in Australia, social security/welfare business and data consist of comprehensive characteristics and complexities specific to the data mining community which are not comparable with those in many domains. This is reflected through the nature of mixing politics, economy, society, organizational and business processes, and the Internet, as well

[3] One at the University of North Carolina (UNC group) (http://ssw.unc.edu/ma/index.html) and the other is our group (UTS group) (http://datamining.it.uts.edu.au/ssdm).

as the distribution of information sources, business dynamics, customer–government interaction evolution, and integration of business and technical issues. This makes social security/welfare data very challenging and complex to analyze.

3) In fact, social security data/business provides a relatively complete testbed for both existing and emerging research on data mining thanks to the complexities from data to problem modeling and delivery of knowledge. The mixture of several complex aspects makes SSDM even more challenging, as, for instance, in mining impact-oriented behavior patterns in large-scale of data mixing customer demographics, activities, policies, and performance.
4) Specifically, very limited SSDM work has been done in either research or practice, leaving a big gap in relation to increasing business needs. Besides the data/business characteristics common to many other areas, it is worthwhile and highly demanding to explore characteristics and challenges in the marriage of data mining with social welfare business and to systematically explore business problems, research issues, challenges, limitations in directly applying existing data mining outcomes, and opportunities to invent new techniques.
5) While SSDM involves many challenges common to other domains, the mixture of specific business mechanisms with wide data complexities also makes SSDM important and challenging, in aspects such as specifying/customizing and inventing data mining methods and algorithms to effectively analyze social welfare business, e.g., processing specific data characteristics and discovering patterns therein.

Since 2004, we has been engaged in data mining for social welfare business. So far, several projects have been conducted which tackle issues, such as analyzing customer earnings [126], profiling debt recovery and verifying changes in earnings declarations [127], investigating relationships between activity sequences and debt occurrences [24], modeling activity impact on debt risk and cost [23], [112], [114], identifying high impact activities/activity sequences on debt occurrences [23], rating customer risk on causing debt [128], fraud detection [129], and so on. Section V will present more details and references about our SSDM-related practices in Australia.

Starting from the understanding of social security business and data, the following sections present an SSDM framework and address SSDM goals, tasks, and challenges. We also summarize the real-life practices of SSDM in Australia and discuss the extension of SSDM for mining general public-sector data.

III. Social Security Services and Data

A. Social Security Business

In countries like Australia and Canada, a variety of social welfare allowances/services and social programs are provided by the government to assist people to become self-sufficient and to support those in need. Fig. 1 illustrates a cause–effect relationship between customers and government in the social welfare business. A customer lodges an allowance application, which is checked by the government. Payments are arranged on the premise of customer entitlement and policies. The customer is required to declare any changes that may affect payment entitlement. Once a customer declaration is lodged, it will be verified by the government. As a result, customer payments are further verified and adjusted if necessary. In some cases, overpayments to the customer may occur for reasons such as incorrect declaration. The government will seek to recover the debt, and the customer will be requested to pay back such overpayments through repayment arrangements made between the government and the customer. More information about social welfare business can be found on the respective government websites[4] and in reports [121].

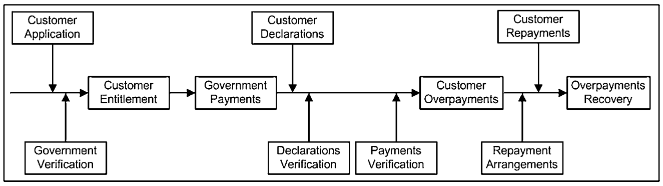

Fig. 1. Social security business workflow.

Taking the Australian social welfare business as an example, as a one-stop-shop, the Australian Commonwealth Department of Human Services (Centrelink[5]) delivers a range of Commonwealth services to the Australian community and is responsible for the distribution of around $86.8 billion, i.e., 30% of the Commonwealth's outlay, to about one-third of Australians [121]. In day-to-day interactions between the government and customers, Centrelink accumulates a large amount of interaction data. For instance, Centrelink provides 361 000 face-to-face services each working day and processes 6.6 billion transactions against customer records each year [122]. It has been shown that the number of such interactions increases every year. The government has progressively recognized the importance of analyzing these interactions to obtain a deep understanding of customers and organization–customer relationships, to actively manage customers, to improve government service quality and objectives, and to inform policy design.

An issue of particular interest is the identification of the drivers that cause noncompliance in organization–customer interactions. Noncompliance drivers may result from many aspects or staff errors. The 2007–2008 audit report by the Australian National Audit Office (ANAO) [119] drew attention to the importance of deeply understanding customers, and of addressing the behavior and behavioral changes in rising debt from the perspective of the customer, government administration, client group, and community.

B. Social Security Data

As a result of implementing day-to-day social security services, a huge amount of data has been accumulated which

[4] Human Resources and Skills Development Canada: http://www.rhdcc-hrsdc.gc.ca/eng/home.shtml; Australian Commonwealth Department of Human Services: http://www.humanservices.gov.au/corporate/home

[5] www.centrelink.gov.au

TABLE 1
CENTRELINK BUSINESS DIMENSIONS 2008–2009

Dimensions	2008-09
Payment value made on behalf of policy departments	$86.8 billion
Debt raised	$1.926 billion
Number of debts	2 187 821
Customers	6.84 million
Individual entitlements	10.43 million
New claims granted	2.7 million
Phone calls	33.7 million
Letters to customers	109.5 million
Online transactions (online and view)	over 24 million
Transactions on customer records	over 6 billion
Mainframe disk capability	550+ terabytes
Eligibility and entitlement reviews	3 867 135
Service delivery points	more than 1000
Customer service centres	316
Centrelink agents and access points	568

increases dramatically every day. Table I[6] provides an overview of some Centrelink business dimensions related to data [121].

As Table I shows, the huge amount of social security data accumulated by Centrelink consist of very useful information recorded from customer service centers, agents and access points, the Internet, interviews and reviews for all services, customers, staff and agents, and debt. The $1926 million in debt raised in 2008–2009 compares with $1831 million in 2007–2008. Such data can be classified into the following categories:

1) Customer demographic and circumstance data, recording information about a customer and his/her circumstances, circumstance changes, etc; for instance, home address and the history of address change;
2) Benefit/allowance data, regarding the information about specific benefit/allowance design and applicability in alignment with customer eligibility, and management processes;
3) Customer pathway data, reflecting the history and relevant details of a customer's use of government services, such as the number of services, when, and from which service centers the services have been applied for, and granted;
4) Activity data, providing activity records information about who (maybe multiple operators) processes what types of activities (say change of address) from where (say customer service centers) and for what reasons (say the action of receipt of source documents) at what time (date and time), as well as the resultant outcomes (say raising or recovering debt) [24], [120];
5) Facility usage data, regarding the resources used by or for customers, e.g., phone calls and online services;
6) Service policy data, information about policies, the applications of policies to customers with particular circumstances;
7) Service transactional data, day-to-day information recorded regarding the use of services, such as new registrations, new claims, debt review, etc.;
8) Service performance data, concerning service quality and performance, such as overpayments and their distribution, how long on average a customer has to queue, general customer satisfaction, etc.;
9) Interaction data about communications between customers and staff, and between staff from different units; for instance, a customer calls Centrelink to report an income update;
10) Operation data about the resources and infrastructure used for day-to-day business, e.g., how many staff hours are spent on payment reviews;
11) Operational performance, concerning the performance of operational expenses and resource use; for instance, the average cost of recovery per dollar of debt, or the effectiveness of reminder letters in terms of solving problems (such as seeking to recover the outlays).

From the aforementioned summary, we identify social security data as having the following characteristics.

1) *Large-scale:* as shown in Table I, huge amounts of data are collected every day and every year.
2) *Mixed structure:* Data incurred by the business consist of all major types, such as numerical, categorical, textual, discrete, continuous, temporal, and sequential.
3) *Distribution:* Data are collected from service centers, the internet, and access points, and recorded in mainframe storage distributed in large centers; customers are distributed everywhere across the country.
4) *Longitude:* Typically, a customer is engaged with a service for quite a while before they are terminated or transferred to another service type.
5) *High dimension:* Data involve multiple dimensions; for instance, there are over 200 types of activity codes reflecting different actions taken in customer–Centrelink interactions.
6) *Multisource:* Different aspects of information data are recorded separately; any one source of data is insufficient to generate a full picture of a particular service.
7) *Sparseness:* The longitudinal data are generated on demand, which is normally random and infrequent; the resulting data are very sparse; for instance, a customer may have accessed a service center two years previously and, later, contacted another office in another place to update a circumstance change, or request a new service.
8) *Imbalance:* Data are not equally distributed, with some being of much higher frequency than others; for instance, outlays only consist of a very small proportion of the overall expenses in Centrelink; customer–government interaction data are not equally distributed, and some activities occur much more frequently, or in more places, than others; Debt-related data only constitute a very small proportion of social security data [23], which forms a class imbalance. In addition, the customer–government interaction activities that are related to debt are composed of a very limited portion of the whole activity set, which gives rise to an item-set imbalance issue.
9) *Divided quality:* It is known that, with data being recorded in divided quality, some data may be missing, or recorded in duplicate.

[6] Data collected from the Centrelink Annual Report 2008–2009.

10) *Variation:* Changes happen everywhere, involving all of the above aspects and data characteristics; in fact, as identified in the ANAO report [119], changes have a critical effect on business integrity and performance stability.
11) *Coupling relationship:* Data entities (objects) and values are often inter-related because of intrinsic business logics in social security. For instance, a debt may be incurred as a result of a wrong declaration of income and address change; a follow-up arrangement activity is made for the customer to pay back the debt once confirmed (called repayment); the customer may either follow the arrangement (refers to the activities arranged by the government) or take other actions to address the repayment. This example shows that objects (i.e., customer, debt, arrangement, and repayment), transactions (regarding customer, debt, arrangements, repayment, etc.), and behaviors from different objects are inter-related.

The aforementioned characteristics are typically aligned with data complexities currently explored in the broad data mining community. They also create additional challenges for existing data mining methods and algorithms when they are deeply engaged in the social security business. In fact, the social security area presents great possibilities to explore typical data mining complexities in one domain. This leads to the need for further research on SSDM tasks and challenges (see Sections IV-D and IV-C).

IV. Framework of Social Security Data Mining

A. Basic Framework

Like any other domain, data mining applications in social security are driven by business objectives and underlying data. Based on the introduction of social security business and data in Section III, Fig. 2 presents a high-level SSDM framework. It consists of three layers: the data layer, the business objective layer, and the data mining goal layer.

The business objective layer includes the main aims and expectations for the implementation of social security services. For instance, Fig. 2 lists the main objectives [121], including customer service enhancement (to instantly provide high-quality services to those with particular needs), payment correctness enhancement (e.g., to pay the right amount to those who are eligible), business integrity enhancement (e.g., to improve the consistency and accuracy and to speed up processing), debt management and prevention (e.g., to recover and prevent debt instantly), outlays cause identification (e.g., to identify outlays incurred by staff error), income transparency improvement (e.g., to improve customer earnings reporting and to detect gray income automatically), performance enhancement (e.g., to reduce customer waiting time in service centers or call centers), service delivery enhancement (e.g., to strip out unnecessary contacts and provide easier and more efficient pathways to services), service/risk profiling (e.g., to identify customers most at risk of incorrect payments and to identify opportunities to reduce the debt more efficiently), customer need satisfaction (e.g., to identify customers with special or more urgent needs than others), accountability assurance (e.g., to identify areas of significant financial or operational risk and to pinpoint more effective arrangements to manage risks), fraud and noncompliance detection (e.g., to identify international or staff fraud and noncompliance), process optimization (e.g., to streamline processes for easier service access and delivery), and key performance indicator (KPI) enhancement (i.e., to identify where and how the key performance indicators can be enhanced).

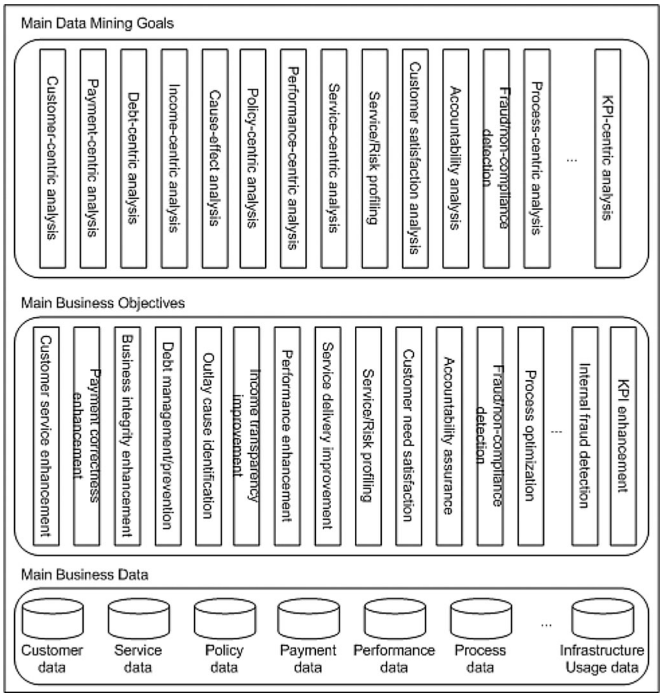

Fig. 2. SSDM framework.

To support the aforementioned major business objectives, the government invests in efficient information infrastructure. As a result, data are acquired and constantly updated at every place and time in the business operation. The data layer summarizes the main data resources. It consists of customer data (customer demographic and circumstance information), service data (service usage and procedural information), policy data (government policy and the applications of policy), payment data (customer payment information), performance data (service performance and operational performance), process data (business process and change applied to customers), infrastructure usage data (the use of IT resources and services), etc.

While every effort has been made to rectify problems, it has been disclosed that the government is facing longstanding, as well as emerging, problems in achieving and improving the main business objectives [119]. The accumulation of business data provide a unique and essential premise to disclose hidden and implicit channels, indicators, and solutions for these issues, as shown by the data mining pilots in Centrelink (see Section V for more information). The data mining layer lists the main goals in mining social security data to enhance business objectives; for instance, customer-centric analysis, payment-centric analysis, debt-centric analysis, income-centric analysis, cause–effect analysis, policy-centric analysis, performance-centric analysis,

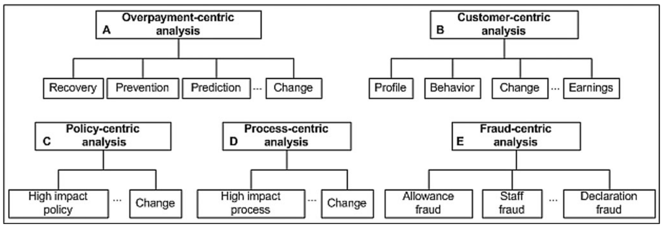

Fig. 3. SSDM goals.

service-centric analysis, service/risk profiling, customer satisfaction analysis, accountability analysis, fraud/noncompliance detection, process-centric analysis, and KPI-centric analysis. These processes will be discussed further in the following sections.

B. Social Security Data Mining Goals

We summarize the main data mining goals in the social security area into the following five classes, according to our understanding and practices of key entities, problems, and challenges in social welfare business by data mining: 1) overpayment-centric analysis; 2) customer-centric analysis; 3) policy-centric analysis; 4) process-centric analysis; and 5) fraud-centric analysis [121]. They are shown in Fig. 3 and are explained briefly in the following.

1) Payment-Centric Analysis: Overpayments or government customer debt are a major concern in social security government services [119]. Overpayment/debt-centric analysis, therefore, emerges as a major objective of SSDM. Its goals consist of the deep understanding of the distributions of overpayments across business lines, the cause and effect of overpayments, and the evolution and changes of overpayments in the life of government customers. In addition, issues that are related to payment accuracy also involve underpayment analysis, and alignment and gap analysis between customer earning/employer payment and government payment. The findings from payment-centric analysis contribute to government customer debt recovery, debt prevention, and debt prediction, as well as better customer service quality.

2) Customer-Centric Analysis: Customer-centric analysis in SSDM aims to deeply understand which customers cause overpayments, and the reasons and indicators behind those customers who owe the government [119]. The reasons may be related to customer profiles, behaviors, earnings, and so on, as well as changes to any of these aspects. The findings from customer-centric analysis contribute to evidence, indicators, and observations that assist the understanding of why, when, and how some customers cause overpayments when others do not. In addition, customer risk rating and customer service recommendations are other objectives.

3) Policy-Centric Analysis: Policy-centric analysis in SSDM seeks to deeply understand which policies [121] are associated with overpayments, and the reasons and indicators behind these policies. Other analysis may focus on the relationships between policy changes, overpayments, and customers. Identifying those policies and policy changes that have led to, or are associated with, overpayments could be used to prevent the occurrence of debts and to actively manage customers.

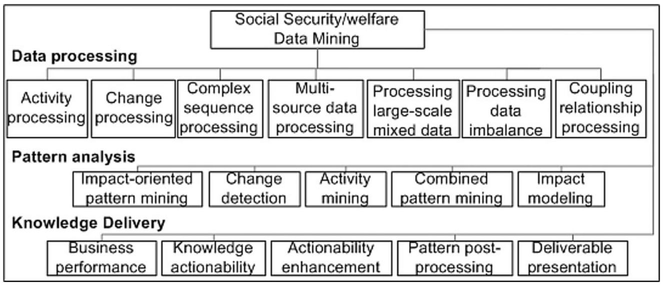

Fig. 4. SSDM challenges.

4) Process-Centric Analysis: Process-centric analysis in SSDM is carried out to deeply understand what business processes or process changes are associated with overpayments [119], [121], as well as the reasons and indicators behind them. By analyzing the relationships between processes, overpayments, and customers, social security government officers obtain a deep understanding of what could be optimized in business processes or during process changes in order to minimize overpayments or the probability of debt occurrence.

5) Fraud-Centric Analysis: Fraud-centric analysis in SSDM is undertaken to analyze whether fraud takes place in the social security business, and where, why, and how fraud happens and evolves [133]. Analysis can be conducted on child welfare fraud, allowance fraud, declaration fraud and staff fraud, and the resultant findings that are used to assist the detection, prevention, and prediction of fraud in the social security business.

C. Social Security Data Mining Challenges

The data mining tasks listed in Section IV-D are comprehensive, involving many aspects of both traditional and emerging data mining techniques. Some can be handled by the utilization of existing general data mining techniques, while others have to be dealt with using revised or newly developed methodology and approaches in order to handle the mixture of social security business and general data complexities. On the basis of our observations of the challenges involved in conducting the aforementioned data mining tasks, we discuss the following key challenges (see Fig. 4) in terms of the main procedures of social security data processing, pattern analysis, and knowledge delivery.

1) Social Security Data Processing: The processing of data characteristics in social security business shares many data-processing issues within the data mining and machine-learning community. In particular, the following areas are especially, important in SSDM.

a) *Activity processing:* The processing of activities and activity sequences [24] needs to address complex features, such as temporal, spatial, structural and semantic dimensions, as well as handle issues such as data sparsity and imbalance [23], [107], dynamics, and associated impact [25] on business (such as causing overpayments) in activity feature selection, activity extraction, activity sequence construction, and preparation.

b) *Change processing:* Changes are widely dispersed in social security data [119], [131]. The consideration of change data in SSDM is crucial to identify meaningful causes and patterns [19]. Change data processing involves issues such as change definition, representation of change, interactions and relations between changes and other entities, and change feature extraction.

c) *Complex sequence processing:* Customer–government interaction generates intensive activities in the social security business. Sequences are complicated [22]–[24] if the relevant information is considered; for instance, the time an activity takes place and the reason (it could be another activity) for the activity occurrence. For family-based debt investigation, it is probably necessary to put all family members' activity sequences together to observe the differences, which will involve multiple coupled sequences [19]. The processing of such sequences involves issues such as representation of single and multiple coupled sequences, modeling sequence relations, sequence feature extraction, and data structure design for storing sequences, and relevant information.

d) *Multisource data processing:* Very often, SSDM has to engage multiple sources of social security data, since more informative patterns reflecting the actual business picture can only be identified on multisource data [22], [110]. Meaningful SSDM analysis involves data of customer–officer interaction transactions, customer demographics, government policies, business processes, customer registration data, customer earnings, debt outcomes, and debt recovery arrangements [121]. It is necessary to correctly understand the relationship between different sources of data from business logic, syntactic and semantic aspects, how to align and fuse them (for instance, whether from the data or pattern [22], [112] perspective) while considering the intrinsic business logic, and how to deal with different granularities.

e) *Processing large-scale mixed data:* SSDM tasks often involve large-scale mixed data. For instance, a debt usually occurs and exists for several months to a few years [130], and the investigation of debt drivers needs to involve multisource information recorded in different structures and formats [121]. Among other things, this requires determining the timeline to select and align different sources of data [22], [110], a smart data structure to fit relevant information, proper sampling methods, an efficient strategy to scan/filter the data, and the selection and fusion of mixed features in both processing and pattern mining [113].

f) *Processing data imbalance:* Besides the normal techniques available in the literature about class imbalance [23], [107], such as undersampling and oversampling, it is expected that greater effort will be necessary in processing extremely imbalanced data in a large scale set and on designing proper data structures, filtering, and sampling algorithms to prepare both class and item-set imbalanced data.

g) *Coupling relationship:* The representation of coupling relationships [18], [19], first involves the definition and extraction of coupling within an entity and between entities from syntactic and semantic perspectives. Further work concerns how to manage the relationships and store the relevant data by developing a suitable data structure, a representation system, and data extraction mechanism.

2) Social Security Pattern Analysis: While many traditional and emerging pattern mining methods and pattern types can be applied directly to SSDM, we also observe specific needs emerging from mining social security data. These are briefly discussed below, and the observations can certainly be used for mining other, similar applications.

a) *Change detection:* Challenging issues in detecting changes [131] cover many possibilities, e.g., representing changes and change contexts in organization–customer interactions, tackling data complexities in processing and mining change-centered data, identifying change patterns in customer circumstance and behavior contexts, identifying group relationships and group behavior changes, identifying customer interaction changes in response to policy/procedure changes, adapting the detection of customer and group dynamics, and extracting and evaluating noncompliance drivers based on the mined change patterns.

b) *Activity mining:* Activities [23], [24] are widely seen in social security data, which create a challenge for traditional data mining [24]. Mining activity patterns can focus on activity-centric, impact-centric, or customer-centric analysis [24], and each aspect is new. In addition, the evaluation of activity mining is nonexistent, and therefore, new interestingness metrics need to be developed for each activity mining method.

c) Impact modeling measures the impact of certain data on business [23], [116]. The nature of the impact, how to measure it and how it is associated with patterns and debt occurrence, is open to investigation. The measure of impact needs to be specified in terms of target data and customer groups by involving domain knowledge and needs to be evaluated by domain experts.

d) Impact-oriented pattern mining identifies patterns that are associated with specific impact. Unlike traditional patterns that consist of items only, impact-oriented patterns have two facets: one is item sets, the other is the impact associated with the item sets. As discussed in [23], impact-oriented pattern mining is challenging, since many emerging pattern types may be identified, such as positive-impact-oriented patterns, negative-impact-oriented patterns, impact-contrasted patterns, and impact-reversed patterns [22], [23].

e) Combined mining [22], [105], [110], [112] is a two to multistep data mining procedure, consisting of mining atomic patterns, merging atomic pattern sets into a combined pattern set, or merging dataset-specific combined patterns into the higher level of a combined pattern set if there are multiple datasets. Combined mining is essential in SSDM, which mines for combined patterns by engaging multisource data and complicated social security data, as discussed in the previous section on data processing. Mining combined patterns is not easy and involves the invention of new techniques and methods. For instance, the authors in [26] have discussed new methodologies, including multifeature combined mining, multimethod combined mining, and multisource combined mining [22]. Combined mining can lead to creative pattern types, such as pair pattern, cluster pattern, incremental pair pattern, and incremental cluster pattern [22], in which pattern components are coupled in terms of relationships such as peer-to-peer or master–slave.

3) Knowledge delivery: In actionable knowledge discovery [17], [20], [26], it is important to deliver knowledge of business interest which can be taken over by business people. This is not a trivial problem, as discussed in domain-driven data mining [7],[7] and is applicable to SSDM.

a) Business performance is the performance of patterns from the business perspective [17]. While data miners usually concentrate on the technical performance evaluation of patterns, the specification and evaluation of business performance will certainly provide additional information for business people to judge the value of the findings. How to define and measure business performance from subjective and objective perspectives is worthy of research, as well as which business metrics need to be defined for use to generally measure business impacts associated with patterns [20].

b) Knowledge actionability is the actionability of identified patterns. The authors in [17] propose a general framework of knowledge actionability and highlight the engagement of both technical and business performance from subjective and objective perspectives in measuring knowledge actionability. In the social security area, our job is to specify metrics to measure SSDM knowledge actionability.

c) Actionability enhancement [17], [20] concerns enhancing the actionability of identified patterns. While many aspects can be addressed [17], it is not a straightforward task. It is worthwhile to analyze why the discovered knowledge is not actionable, what aspects can be focused on, and what actions can be taken to enhance the actionability.

d) Pattern postprocessing is an important way to enhance knowledge actionability and is applicable to SSDM. The authors in [113] summarize the main techniques to be developed or enhanced in postprocessing and postmining and collect the latest work on post mining of association rules. Considering the social security data specialization, new postprocessing techniques need to be developed, with the involvement of domain knowledge.

e) Deliverable representation [17] builds appropriate mechanisms to convert SSDM findings into business-oriented deliverables and represents deliverables in a business friendly manner. This is actually a challenging issue, which has not been well studied. Because of the engagement of customer–government interaction, in particular, the relevant deliverables need to show the interactive procedures by reflecting the underlying activity patterns.

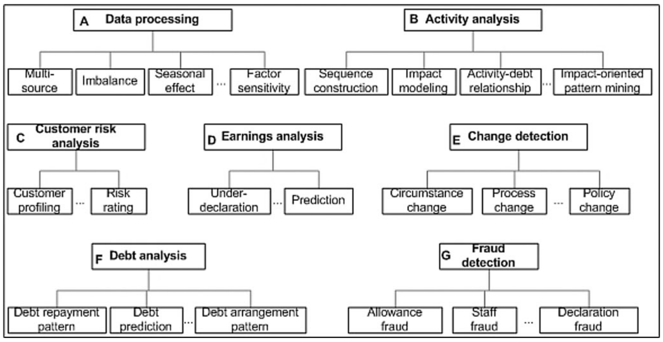

Fig. 5. SSDM tasks.

In Section V, we introduce a number of case studies of SSDM that address some of the aforementioned challenges.

D. Social Security Data Mining Tasks

To support the data mining goals that are discussed in Section IV-B, many tasks need to be performed in SSDM. Traditional data mining methods, including association rule mining, frequent pattern mining, clustering, and classification, will certainly play an important role in achieving the aforementioned goals. We categorize the SSDM tasks as follows, according to the main entities in social security/welfare business [121] and to address the SSDM goals, as shown in Fig. 5: 1) data processing; 2) activity analysis; 3) customer risk analysis; 4) earnings analysis; 5) change detection; 6) debt analysis; and 7) fraud detection. We briefly explain these tasks in the following.

1) Data Processing: The main tasks in social security data processing have many aspects, including many common data-processing issues discussed in the community. In particular, we have the following.

a) SSDM involves multiple data sources: for instance, customer demographic data, customer interaction transactions with government officers, arrangement and repayment activity data, and debt-related data. Therefore, it is essential to deal with multiple sources of data [22].

b) *Imbalance:* Overpayment-related data only consist of a very small proportion of the whole social security data. Debt-related pattern analysis has to identify patterns in imbalanced datasets [23].

c) *Seasonal effect:* Social security government services present very strong seasonal characteristics that are determined by service objectives and policies. For instance, during holiday seasons, many immigrants move to their mother country, taking children to visit relatives. This may lead to gaps in reporting, although the government may continue to pay the usual rate, resulting in a government overpayment [130].

d) *Factor sensitivity:* This reflects the fact that not all variables and values contribute equally; some variables may play only a small role or duplicate others. Factor impact analysis, principal component analysis, and feature mining

[7] DDDM-SIG: http://datamining.it.uts.edu.au/dddm/

may be necessary in analyzing the sensitivity and interrelationships amongst factors, variables, and features.

2) Activity Analysis: Activity data [24], [120] refer to the interactive events, operations, and actions occurring in social security business. They form the main component of behavioral data [15], and the analysis of activities is complicated [24]. We discuss several tasks here.

a) *Activity sequence construction:* This involves activity types, activity distribution, activity relationships, timeline, and so on. The exploration of these aspects can generate useful hints for the construction of activity sequences [23], [120].

b) *Activity impact modeling:* In business, each activity or activity series plays a different role, and some activities contribute more than others. Different combinations of activity sequences may lead to a variety of outcomes. Before constructing activity sequences, there is a need to understand and quantify the outcomes and the impact of activities associated with a particular business, e.g., debt occurrence. Measures and models need to be developed to specify and differentiate the impact of particular activities [23], [24].

c) *Activity–debt relationship analysis:* In the social security domain, the occurrence of debt is largely driven by activities or activity sequences. The analysis of relationships between activities and debt [23], [24] aims to determine which activities are more sensitive to debt occurrence, how they result in debt (e.g., as a group, or before or after the debt occurrence), and to what extent an activity (sequence) leads to debt.

d) *Impact-oriented pattern analysis:* While general activity patterns can be identified, business people are more interested in those activities that are associated with high business impacts, which we call *impact-oriented patterns* [23], [24]. Mining high business impact-oriented patterns is not easy, since it may involve the handling of activity imbalance, impact definition, complex pattern types, and the definition of new interestingness metrics.

3) Customer Risk Analysis: As a result of having a deep understanding of customers, any customer or customer group can be ranked in terms of the risk of causing overpayments. To rate customers requires consideration of various scenarios and risk specifications.

a) *Customer profiling [130]:* This creates a comprehensive understanding of which customer profiles lead to debt at different probability levels, and which profile-based factors are more sensitive to which allowance-based debt occurrence. Customer profiling needs to be more deeply conducted by scrutinizing customer circumstances, distributions, structures, relationships, and their variations. It is worthwhile to analyze the relationships between these aspects and debt occurrence and debt impact.

b) *Customer risk rating [126]:* While it is known that some customers are more likely to be associated with debt occurrence than others, it is advantageous to specify their particular risk and risk rate. This is associated with issues, such as risk types, which customer-related factors contribute to risk from the perspective of demographics, behaviors and change, and time sensitivity to risk occurrences. It is also interesting to see whether some customers have a higher probability than others of causing debt before they register an allowance, and the key factors causing such a difference. Information declared by a customer to the government certainly affects the likelihood of risk and debt. We are interested in the relationship between the information declaration level and coverage and the risk level.

4) Earnings Analysis: Incomes and earnings are particularly sensitive to social security debt and the delivery of service objectives. Any deliberate manipulation or unintentional disregard of earnings could lead to eventual overpayments. Therefore, earnings analysis not only concerns the relationship between what has been declared and the debt occurrence, but also what has been under-declared or is missing.

a) *Manipulative declaration analysis:* The manipulation of earnings declaration has been viewed as one of the key drivers of debt [127], [132]. Earnings may be underdeclared or neglected in reporting to the government, and the direct detection of manipulative declaration is often difficult, since it is not easy to identify the evidence. The identification of manipulative declaration needs to capture what a customer reports to the government at the initial registration, customer behavioral data, customer circumstance changes, family-based behaviors and situation changes, and customer activities related to expenses, which is often a very complex issue.

b) *Earnings prediction [132]:* While it is difficult to achieve, the prediction of earnings for correctly and manipulatively declared customers who are eligible for government benefits can assist with the early detection, and thus prevention, of debt. The detection of earnings for correctly declared customers is generally more manageable than for manipulators. Besides numerical data-based prediction techniques, new prediction methods are essential by combinatorially considering customer circumstance changes, behaviors, membership data and changes, etc.

5) Change Detection: In organization–customer interactions, significant changes, occurring either in customer circumstances and behaviors, or in business policies and processes, may lead to noncompliance, resulting in inconsistencies or even substantial financial losses and damaging effects on an organization [131]. For instance, changes in customer demographics may not be instantly reflected in relevant business lines, thereby resulting in inconsistencies or overpayments. An example is the almost $2 billion Centrelink customer debt, which the 2007–2008 ANAO audit report [119] concludes that customer debt arose primarily from customers failing to notify Centrelink of changes in circumstances. This is a challenging issue.

a) *Customer circumstance change analysis:* Customers often experience change in their circumstances, e.g., changing their home address or educational status. Any significant circumstance change could lead to debt, and the detection and early prediction of such significant circumstance changes are critical for managing debt. In addition, changes that are associated with specific customer groups are interesting. Identifying such groups and their behavior and circumstance changes is useful in understanding the reasons for debt occurrence.

b) *Policy change analysis:* Some policy changes have been shown to be related to debt occurrence. A policy change may affect certain customers' behaviors and declarations and eventually lead to more debt. It is worthwhile to analyze the relationships between particular policies (or policy groups), customer behavior changes, declaration changes, and debt occurrence. Key issues include the analysis of which policy changes are most likely to cause debt, and why that happens. The findings can then be used to advise policy changes and to take steps toward customer intervention for certain policy changes.

c) *Process change analysis:* Similar to policy changes, some process changes are more sensitive to debt increases than others. The analysis of relationships between process changes and debt changes can alert process makers to intervene in certain process changes to prevent associated debt. Similarly, it is helpful to combinatorially analyze process changes, customer behavior changes, declaration changes, and debt changes.

6) Payment Accuracy Analysis: As the actual outcome of unfair and wrongly directed social security delivery, the payment and debt-centric analysis is designed to discover direct characteristics, distributions, causes, and changes associated with debt occurrence. Many aspects can be studied in payment and debt analysis; for instance, debt statistics to generate the distribution of debt and the dynamics of debt development.

a) Debt recovery pattern analysis [103], [115] is conducted to analyze patterns of debt-recovery-related activities. This can be conducted on the recovery activity sequences, time interval distributions for different recoverable groups, recovery speed analysis, recoverable customer characteristics, unrecoverable customer circumstance patterns, effectiveness of recovery methods on different customer groups, and early recovery recommendation.

b) Debt arrangement pattern analysis [23] analyzes patterns of government arrangement-related activities and methods. It is worthwhile to analyze the arrangement effectiveness of arrangement methods and intervals in relation to different customers or customer groups. With the findings of this research, more pertinent arrangement methods and time arrangements can be made for particular groups.

c) Debt repayment pattern analysis [23] is carried out to analyze patterns of debt-repayment-related activities and repayment groups. This can be through customer repayment activity sequences, repayment time interval distributions, repayment method distribution, fast–slow repayment characteristics, effectiveness of different repayment methods on different groups, and effective repayment recommendations to particular groups.

d) *Debt arrangement–repayment analysis [23]:* By combining the analysis of debt recovery arrangements and customer repayments, more actionable patterns can be identified to advise the effective arrangement–repayment combinations for particular customer groups to enable more effective recovery. The combinatorial analysis of customer circumstances, arrangement methods, repayment methods, intervals, debt recovery speed, etc., can lead to very informative intervention rules for debt recovery.

e) Debt prediction [103], [108], [115] aims to predict which customers or customer groups will incur a debt, on which benefit services, and when. By focusing on particular customers or customer groups, the task is to predict when, or at what time interval, a debt will occur, the likely frequency of debt occurrence, the likely size of the debt, and so on.

f) Driver analysis is for customers with either overpayment or underpayment, or for those with misaligned payment between earnings and entitlement. Detection of drivers may be conducted from many aspects, such as the differences between over- and underpayment groups, changes associated with normal and incorrect payment groups, and behavioral patterns associated with earnings declaration of those customers whose income is misaligned with the entitlement.

7) Fraud Detection: Fraud may take place in many aspects of the business.

a) Allowance fraud [121], [133] is that which takes place on allowances. There are many types of allowances that are available for eligible customers from government services. A customer cheating the government to obtain an allowance to which they are not entitled causes an allowance–customer mismatch fraud. In other cases, an eligible customer may cheat the government in order to maximize the amount of payment, resulting in overclaim allowance fraud. Allowance fraud detection aims to detect the corresponding key factors contributing to allowance–customer mismatch frauds and overclaim allowance fraud, and the patterns and reasons that relate to them.

b) Declaration fraud [121], [133] is the fraud that takes place on declarations, which may be manipulated by fraudsters. Declaration fraud detection identifies factors, scenarios, patterns, and changes of underdeclaration, delayed declaration, and missing declaration in terms of allowance types and customer groups. Different allowance types and customer groups may experience a variety of declaration patterns. Other work includes the prediction of manipulative declarations.

c) Staff fraud [121], [133] is the acquisition of payments by an employee, for himself or herself or for another person, through dishonesty or deception. A staff member may create false documents and process false benefit claims against genuine customer records. Staff fraud detection seeks to identify key factors, business sections, methods, and patterns related to the fraudulent behavior of staff and the impact of such behavior. It is often difficult to identify staff fraud because it involves complex data, business process, and lack of evidence.

V. Case Studies of Social Security Data Mining Practices

A. *Australian Practices*

Australia is one of the most advanced countries in terms of inventing SSDM to enhance decision support systems. Currently, the main techniques and methods for decision support consist of data matching, service profiling, business intelligence, and data analytics.

1) Data matching refers to the process and tools to match social security data against other sources of data, such as from immigration, customs, taxation, and banking systems.
2) Service profiling profiles customers associated with different services.
3) Business intelligence usually refers to data warehousing and reporting, including *ad-hoc* and online analytical processing-based analysis.
4) Data analytics covers a broad scope from descriptive analysis to data-driven pattern and anomaly detection and analysis.

Since 2004, we have been engaged with Centrelink (now the Department of Human Services) in conducting data mining on social security data[8] [114]. A series of research and commercial projects have been conducted to explore key factors and variables [105], [107], [126], behavior and interaction patterns [23], [24], [108], [130], abnormal circumstance changes [111], [131] and earning/income declarations [103], [111], [115], [127], [132], customer levels of risk relevant to overpayments [103], [106], [108], [115], [130], fraud detection [129], and risk management for online income declarations [132]. This has led to several algorithms and models being specifically designed to analyze social security data; for instance, high impact-oriented activity pattern mining [23], positive and negative sequential patterns (NSPs) of debtors and nondebtors [111], [116], behavior sequence classification [105], [106], [108], and combined pattern mining [26], [106], [112]. The patterns and results delivered in Centrelink projects involve many millions of dollars in the recovery and prevention of overpayments for the government. To the best of our knowledge, there is no similar substantial work conducted on mining social security data in the community.

These projects address many business concerns, as discussed in Section IV-B, and data mining tasks, as outlined in Section IV-D. Through these preliminary studies, we accumulate a fundamental understanding of how data mining can be appropriately used to address critical issues in the social security domain, the limitations of traditional and emerging data mining techniques, which new techniques need to be invented, and so on. Driven by specific business objectives and processes, social security data present strong concentration and characteristics for data mining and lead to specific research issues that are underdeveloped. This reflects the requirements to be considered when developing SSDM.

In the following, we summarize several SSDM case studies from our real-life projects with Centrelink to address some of the SSDM challenges. Rather than presenting each case in detail, we summarize the cases by highlighting the main business objectives, research issues, and solutions. Where applicable, we refer to relevant documents and papers for detailed techniques and experimental results to enable the presentation of more cases and references to provide readers with comprehensive references and the means to drill down to specific design.

1) Modeling Impact of Activity/Activity Sequence: This is to model the impact (business effect and outcomes) associated with customer activities and customer–officer interactions. For instance, the impact could be government debt, which can be modeled in terms of metrics including customer debt amount and duration, how a debt might incur, as well as the costs for possible debt review against an activity or an activity sequence. Impact is categorized into positive and negative, e.g., debt or nondebt. In [23], models are built to quantify the risk and costs associated with particular activities, activity sequences, and patterns. We defined metrics for this purpose. For instance, *Pattern's Average Debt Amount per Debt* calculates the average amount per debt in terms of the total overpayment for all debts associated with an activity sequence pattern [23]. *Risk of A Pattern* is defined as the ratio of the cost associated with a particular pattern to the total cost of the pattern set identified in the dataset [23].

2) Mining Impact-Targeted Patterns: The analysis of impact-targeted patterns is to identify those patterns with either a positive or negative impact as the target at the right-hand side, in addition to any item sets or sequences at the left-hand side. The identification of impact-targeted sequential patterns affects the construction of sequences and pattern mining with the predetermined association with either positive or negative impact. Cause/effect relationships between activities and their impact (for instance, on debt occurrence) need to be considered. In [23], by combining debt occurrence with activity analysis, we report several algorithms and interestingness metrics to identify different types of impact-targeted sequential patterns, e.g., *impact-oriented pattern*, *impact-contrasted pattern*, and *impact-reversed pattern*. For example, taking debt or not as the impact, a debt-oriented *impact-reversed pattern* P, i.e.,

$$P = \begin{cases} p_1 : \{a_1\} \rightarrow \bar{I} \\ p_2 : \{a_1, a_9\} \rightarrow I \end{cases} \tag{1}$$

consists of two subpatterns p_1 and p_2 with opposite targets (debt I and nondebt $\bar{I}$), in which p_1 is an underlying pattern, while p_2 is derived from p_1 by appending or reducing some activity elements (a_9) which leads to the impact conversion from $\bar{I}$ to I [23], [112].

3) Mining Positive and Negative Sequential Patterns: Based on the presence of activities that both occur and are missing in customer–government interactions, interesting positive (occurring) and negative (nonoccurring) activity sequential patterns [116] can be identified. An NSP, e.g., $\{a_3, \neg a_7, a_{12}\}$ indicates that activities a_3 occurs before a_{12} but without $\neg a_7$ in between. We develop three algorithms to identify NSPs: negative-GSP [118], genetic algorithm-based [117], and e-NSP: a set theory-based approach based on identified positive patterns [36].

[8] datamining.it.uts.edu.au/ssdm

In addition, by involving the pattern impact, algorithms are developed to identify impact-targeted positive and negative sequential patterns, e.g., $\{a_1, a_5, a_{11}\} \rightarrow \bar{I}$. The authors in [111] and [116] discuss definitions, algorithms, and evaluation metrics for impact-targeted NSPs. To evaluate such patterns, new metrics are developed to measure the interestingness, including both the update of existing metrics for positive sequences like *Support*, *Confidence*, and *Lift* [111], and new metrics for the negative sequences such as *Contribution* and *Impact* [116].

4) Mining Attribute Combined Patterns: Attribute/feature-combined patterns [22] refer to a cluster of patterns that consist of attributes typically from multiple data sources, such as customer demographics, customer-government interactions, and debt outcomes. To merge such data is often very costly since a huge number of transactions from different data sources are often involved. We propose the combined mining approach [22], [26], [112], to identify *nonimpact-oriented combined patterns* and *impact-oriented combined patterns*, depending on whether a pattern is associated with a certain target item or business impact. For instance, (1) illustrates a combined pattern pair.

In [23] and [112], we present debt-targeted combined association rules, including rule pairs and clusters. For instance, we find that the combined rule, i.e., $R = \{demographic = \{income_{class} = 0, age > 65\}\}, \wedge\{arrangement = \{withholding\}\}, \{repayment = \{cash|post, withholding\}\} \rightarrow Debt\}$ is highly associated with a low risk of debt (Debt) occurrence (where "demographic" refers to demographic features, "arrangement" comes from the debt payoff arrangement data, and "repayment" means the customer's payback method). For this, we first identify demographic patterns and arrangement–repayment patterns that are associated with debt occurrences in respective data sources and, then, merge/align patterns to filter interesting rule pairs and clusters by evaluating the interestingness of pattern pairs/clusters. Metrics [22], [26], [112] are developed to measure the importance of pattern pairs and clusters. For instance, one way is to select pairs based on the square root of an individual pattern's interestingness if the targets are opposite [22], [112].

5) Identifying High Impact Behavior for Intervention: Customer behaviors and customer–officer interactions are dynamic. In general, not every behavior contributes in the same way. To explore the contribution of those behaviors that play more important roles in causing eventual outcomes, dynamic charting [116] is proposed. The x coordinate grows with the behavior evolution, with each point representing a behavior. The y-axis measures the impact change associated with behavior evolution. Multiple y coordinates may be created to measure the effect or performance of behaviors. Correspondingly, the contribution of each behavior can be measured precisely, and it is easy to identify those discriminative behaviors, for instance, leading the trend change (which we call high impact behavior).

Dynamic charts are very useful for presenting the dynamic evolution of behaviors, the effect of each behavior, and identifying high impact behaviors for early prevention. They can be used to analyze the dynamics of behavioral sequences, activity interaction and evolution, impact change as per behavior evolution, and the formation of associated pairs and clusters in terms of pattern interestingness. With dynamic charts, it is easy to identify those discriminative activities that result in significant variations from their previous adjacent activities. Targeted intervention actions can then be taken to either stop, or prevent, these discriminative activities.

VI. Discussion About Public-Sector Data Mining

The aforementioned case studies and our research work with the Australian Commonwealth department have shown the great potential of applying data mining to social security data, and the opportunities which have arisen to invent new algorithms and tools in data mining. Like other similar areas, they have shown that SSDM can lead to 1) deep understanding of customer behaviors and customer–government interactions, 2) causes and cause–effect relationships of government debt occurrences, 3) relationships between customer demographics, behavior, government policies and arrangements optimization, and customer feedback. The findings have been shown to be useful to inform the government of appropriate strategies from both customer and government perspectives to recover and prevent debt more effectively. Business people are presented with more informative and implicit evidence and indicators for debt prevention and intervention, which are not available from their current systems.

The exercises in the real world also engender lessons that may be helpful for readers who are interested in SSDM research and development, as well as for the general public sector/service data mining [16]. SSDM experience and capabilities are helpful for mining other public-sector data, such as immigration data, custom data, taxation data, child support service data, and medicare data. We summarize some of the lessons and observations that may benefit public-sector data mining.

A. Data Quality and Preparation

1) Generally, public-sector data are of good quality. This does not mean that data cleaning and preprocessing is not necessary to improve data quality for further analysis.
2) In public-sector data, because of the policy/process arrangement and management needs, some variables/attributes may be irrelevant, duplicated or dependent on one another, or present strong seasonal, cyclical or specific benefits or customer group-related effects. It is important to conduct analysis on variables and their relationships and to exclude those variables and data that may affect learning performance.
3) An important goal of mining public-sector data is to identify and investigate problems associated with high business impact, e.g., government debt in taxation and medicare. This leads to the need for impact-targeted analysis [23]. Such impact-oriented data are often imbalanced in terms of impact classes and/or data items, e.g., activities causing debt are often very rare and specific. To cater for such data, different techniques, such as frequency analysis, distribution analysis, downsampling, oversampling, data partition, and metric compensation, may be essential to prepare data for further impact-targeted model design.

4) Because of the nature of government dynamics, public-sector data change from time to time [119]. This results in the frequent change of data distributions and characteristics. For instance, customer–government interaction sequences change with customer or business circumstance changes, while customer reaction activities adapt to policy and process changes. It is important to consider such business change, and the resulting change data and data dynamics. This challenges modeling and necessitates the detection of significant changes associated with particular customer groups or business.

B. Feature and Item-Set Construction

1) Because of the characteristics of public-sector operations, we need to select/construct and mine for discriminative features and variables. Highly dependent and duplicated variables need to be verified and excluded in feature selection and construction.
2) In constructing item sets for pattern mining, many factors need to be considered in order to make the data more meaningful for business. For instance, involving domain knowledge [20], specifying the time window (including the starting point and sliding strategies) for data selection [23], the sensitivity of the selected data against the impact (for instance, government customer debt) occurrence and its effect before and after the impact occurrence.
3) Public-sector data often present distinct political, economic, sociological effects, as well as effects from seasonal, cyclical, regional, and specific benefit or customer group perspectives. Such effects should be considered and analyzed before modeling starts.

C. Model Building

1) While many existing algorithms and models seem to be suitable for SSDM and public-sector data mining, it is necessary to carefully check that they fit perfectly into the public-sector data characteristics. Variations may be necessary. A typical situation is that many service patterns are identified, but most if not all of them simply reflect policy and process arrangements, which is not helpful for government service decision making.
2) While different methods can be used and are helpful, service profiling [121] and domain-driven decision rules [26] are most convenient and effective for public service decision making.
3) Although public-sector data mining shares characteristics and needs with general business applications, public-sector data provide great opportunities to update existing techniques as well as to develop new approaches, in aspects such as customer–government interactions, mixed information from political, economic, sociological, and technical perspectives, from mining single data sources to multiple sources, and from focusing on single business lines to crossing business in relevant service departments. For instance, negative sequence analysis [111], [115] becomes very useful in detecting those customers who intentionally hide some discriminative behaviors. Mining cross-department public-sector data, combined mining [26], [110], [111] is essential.
4) Higher expectation than other business applications is placed on the models that can be self-explanatory to service counter officers, adaptive to policy and circumstance changes [19], operable for people to work on a daily basis, and actionable for results where intervention is necessary.
5) Models developed for public sectors should consider political, economic, and sociological effects, seasonal, cyclical, and regional factors, and should encompass specific benefit or customer group perspectives;
6) Algorithms and models need to be self-adaptive, ideally automatically, to adapt to business and data dynamics. Techniques for building adaptive models and algorithms need to be devised.

D. Evaluation

1) Patterns should be evaluated in terms of the two-way interestingness framework [17], namely against both technical performance and business performance, and from objective and subjective aspects, as well as from political, economic, and sociological aspects.
2) Evaluation metrics to measure the outcome of patterns on business (including political, economic, and sociological aspects) should be specified. In addition, new metrics are necessary for checking group difference and constructing combined patterns.
3) The reliability and security of social security systems could be investigated and enhanced using data mining and machine learning.

VII. Conclusion

With the occurrence of the global financial crisis, more and more governments have realized the necessity of enhancing social security services objectives and quality. Data mining and machine learning can play a critical role, as we have demonstrated in mining Australian social security data for debt prevention, recovery, customer analysis, etc., during the past few years. However, as the literature review shows, mining social security (and public sector) data are still an open field for business applications in data mining and machine learning. Very few references have been publicized. In this paper, for the first time in the community, we present a picture of studies on social security issues and summarize the key concepts, goals, tasks, and challenges of SSDM, based on our experience and knowledge accumulated through conducting data mining in Australian social security data.

We have also highlighted several case studies of mining social security data, including modeling the impact of activity/activity sequences, mining impact-targeted activity patterns, mining positive and negative sequential patterns, conducting impact-targeted sequence classification, and mining combined association rules. We have discussed how the identified patterns are converted into knowledge that can support business people in a more user-friendly way to take decision-making actions.

While these case studies aim to present a picture of what can be done in SSDM, many references have been provided so that readers can access information about the specific techniques in more detail.

We are currently working on the remaining tasks and challenges that are discussed in this paper, such as, detecting fraud in social security data.

Acknowledgment

The authors would like to thank the key staff, including H. Bohlscheid, Dr. Y. Zhao, Dr. H. Zhang, P. Newbigin, and B. Clark from Centrelink, and students and staff who have been engaged in SSDM since 2005, including Prof. C. Zhang, Dr. Y. Ou, S. Wu, J. Li, and Dr. J. Ni, Y. Yang, and Y. Xiao for their collaboration and engagement in the projects.

References

[1] H. Aaron, "Demographic effects on the equity of social security benefits," in *The Economics of Public Services*, M. Feldstein and R. Inman, Eds., London: Macmillan, 2007.

[2] B. Agarwal, "Social security and the family: Coping with seasonality and calamity in rural India," *J. Peasant Stud.*, vol. 17, pp. 341–412, 1990.

[3] L. Alexander and T. Jabine, "Access to social security microdata files for research and statistical purposes," *Soc. Secur. Bull.*, vol. 41, no. 8, pp. 3–17, 1978.

[4] A. J. Auerbach and L. J. Kotlikoff. (1984). "An examination of empirical tests of social security and savings," National Bureau of Economic Research, Cambridge, MA, Working Paper 730. [Online]. Available: http://www.nber.org/papers/w0730.

[5] A. J. Auerbach and L. J. Kotlikoff. (1985, Oct.). "Simulating alternative social security responses to the demographic transition," National Bureau of Economic Research, Cambridge, MA, Working Paper 1308. [Online]. Available: http://ideas.repec.org/p/nbr/nberwo/1308.html.

[6] D. Baker and M. Weisbrot, *Social Security: The Phony Crisis*. Chicago, IL: Univ. of Chicago Press, 2000.

[7] B. D. Bernheim. (1987, May). "Social security benefits: An empirical study of expectations and realizations," National Bureau of Economic Research, Cambridge, MA, Working Paper 2257. [Online]. Available: http://ideas.repec.org/p/nbr/nberwo/2257.html.

[8] R. J. Barro and C. Sahasakul, "Average marginal tax rates from social security and the individual income tax," *J. Bus.*, vol. 59, no. 4, pp. 555–566, 1986.

[9] H. Berghel, "Identity theft, social security numbers, and the web," *Commun. ACM*, vol. 43, no. 2, pp. 17–21, 2000.

[10] D. Blanchet and L.-P. Pele. (1997, Oct.). "Social security and retirement in France," National Bureau of Economic Research, Cambridge, MA, Working Paper 6214. [Online]. Available: http://ideas.repec.org/p/nbr/nberwo/6214.html

[11] D. Bloom, D. Canning, R. Mansfield, and M. J. Moore. (2006). "Demographic change, social security systems, and savings," National Bureau of Economic Research, Cambridge, MA, Working Paper 12621. [Online]. Available: http://econpapers.repec.org/RePEc:nbr:nberwo:12621

[12] R. W. Boadway and D. E. Wildasin, "A median voter model of social security," *Int. Econom. Rev.*, vol. 30, no. 2, pp. 307–328, 1989.

[13] M. J. Boskin and G. F. Break, *The Crisis in Social Security: Problems and Prospects*. Oakland, CA: Inst. Contemporary Stud., 1977.

[14] G. Burtless and R. A. Moffitt, "Social security, earnings tests, and age at retirement," *Public Finance Rev.*, vol. 14, no. 1, pp. 3–27, 1986.

[15] L. Cao, "In-depth behavior understanding and use: The behavior informatics approach," *Inf. Sci.*, 180, no. 17, pp. 3067–3085, 2010.

[16] L. Cao *et al.*, *Social Security Data Mining for Public Services*. [Online]. Available: http://datamining.it.uts.edu.au/icdm10/index.php/case-study

[17] L. Cao, D. Luo, and C. Zhang, "Knowledge actionability: Satisfying technical and business interestingness," *Int. J. Bus. Intell. Data Mining*, vol. 2, no. 4, pp. 496–514, 2007.

[18] L. Cao, Y. Ou, and P. S. Yu. (2011). "Coupled behavior analysis with applications," *IEEE Trans. Knowl. Data Eng.*, to be published.

[19] L. Cao, Y. Ou, P.S. Yu, and G. Wei, "Detecting abnormal coupled sequences and sequence changes in group based manipulative trading behaviors," in *Proc. Knowl. Discovery Data Mining*, 2010, pp. 85–94.

[20] L. Cao, P. S. Yu, C. Zhang, and Y. Zhao, *Domain Driven Data Mining*. New York: Springer-Verlag, 2009.

[21] L. Cao, P. S. Yu, C. Zhang, and H. Zhang, *Data Mining for Business Applications*. New York: Springer-Verlag, 2008.

[22] L. Cao, H. Zhang, Y. Zhao, D. Luo, and C. Zhang, "Combined mining: Discovering informative knowledge in complex data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 3, pp. 699–712, Jun. 2011.

[23] L. Cao, Y. Zhao, and C. Zhang, "Mining impact-targeted activity patterns in imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 8, pp. 1053–1066, Aug. 2008.

[24] L. Cao, Y. Zhao, C. Zhang, and H. Zhang, "Activity mining: From activities to actions," *Int. J. Inf. Technol. Decis. Making (IJITDM)*, vol. 7, no. 2, pp. 259–273, 2008.

[25] L. Cao, Y. Zhao, F. Figueiredo, Y. Ou, and D. Luo, "Mining high impact exceptional behavior patterns," in *Proc. Int. Workshops Emerg. Technol. Knowl. Discovery Data Mining*, 2007, pp. 56–63.

[26] L. Cao, Y. Zhao, H. Zhang, D. Luo, C. Zhang, and E. K. Park, "Flexible frameworks for actionable knowledge discovery," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 9, pp. 1299–1312, Sep. 2009.

[27] R. Carr-Hill, J. Jamison, D. O'Reilly, M. Stevenson, J. Reid, and B. Merriman, "Risk adjustment for hospital use using social security data: Cross sectional small area analysis," *Brit. Med. J.*, vol. 324, p. 390, 2002.

[28] C. Coile, P. Diamond, J. Gruber, and A. Jousten, "Delays in claiming social security benefits," *J. Public Econ.*, vol. 84, no. 3, pp. 357–385, 2002.

[29] H. Cronqvist and R. H. Thaler, "Design choices in privatized social-security systems: Learning from the Swedish experience," *Amer. Econ. Rev.*, vol. 94, no. 2, pp. 424–428, 2004.

[30] P. Cutright, "Political structure, economic development, and national social security programs," *Amer. J. Sociol.*, vol. 70, no. 5, pp. 537–550, 1965.

[31] M. R. Darby, "The effects of social security on income and the capital stock," UCLA Dept. of Econ., Los Angeles, UCLA Econ. Working Paper 095, Mar. 1978.

[32] H. Dean and M. Melrose, "Manageable discord: Fraud and resistance in the social security system," *Soc. Policy Administ.*, vol. 31, no. 2, pp. 103–118, 1997.

[33] P. A. Diamond and P. R. Orszag, *Saving Social Security: A Balanced Approach*. Washington, DC: Brookings Institution, 2005.

[34] P. A. Diamond, *Taxation, Incomplete Markets, and Social Security.* Cambridge, MA: The MIT Press, 2003.

[35] P. A. Diamond, "A framework for social security analysis," *J. Public Econ.*, vol. 8, no. 3, pp. 275–298, 1977.

[36] X. Dong, Z. Zhao, L. Cao, Y. Zhao, C. Zhang, J. Li, W. Wei, and Y. Ou, "e-NSP: Efficient negative sequential pattern mining based on identified positive patterns without database rescanning," in *Proc. Conf. Inf. Knowl. Manage.*, 2011, pp. 825–830.

[37] D. Duncan, H.-C. M. Kum, K. Flair, and W. Wang, "Successfully adopting IT for social welfare program management," in *Proc. Annu. Nat. Conf. Digital Govern. Res.*, 2004, pp. 1–9.

[38] D. Duncan, H. Kum, K. Flair, J. Stewart, E. Weigensberg, and P. Lanier. (2007). NC child welfare program [Online]. Available: http://ssw.unc.edu/ma/index.html

[39] D. F. Duncan, H.-C. Kum, E. C. Weigensberg, K. A. Flair, and C. J. Stewart, "Informing child welfare policy and practice using knowledge discovery and data mining technology via a dynamic web site," *Child Maltreat*, vol. 13, no. 4, pp. 383–391, 2008.

[40] C. A. Echevarría and A. Iza, "Life expectancy, human capital, social security and growth," *J. Public Econ.*, vol. 90, no. 12, pp. 2323–2349, 2006.

[41] Z. Eckstein, M. Eichenbaum, and D. Peled, "Uncertain lifetimes and the welfare enhancing properties of annuity markets and social security," *J. Public Econ.*, vol. 26, no. 3, pp. 303–326, 1985.

[42] M. Gonzalez-Eiras, "Social security as Markov equilibrium in OLG models: A note," *Rev. Econ. Dyn.*, vol. 14, no. 3, pp. 549–552, 2011.

[43] M. Feldstein, "The future of social security pensions in Europe," National Bureau of Economic Research, Cambridge, MA, Working Paper 8487, 2001.

[44] V. Galasso and P. Profeta, "The political economy of social security: A survey," *Eur. J. Political Econ.*, vol. 18, no. 1, pp. 1–29, 2002.

[45] D. M. Garrett, "The effects of differential mortality rates on the progressivity of social security," *Econ. Inquiry*, vol. 33, pp. 457–475, 1995.

[46] C. Gillion, "Social security pensions: development and reform," Geneva, International Labour Office, 2000.

[47] M. S. Gordon, *Social Security Policies in Industrial Countries;A Comparative Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[48] E. M. Gramlich, "Different approaches for dealing with social security," *Amer. Econ. Rev.*, vol. 86, no. 2, pp. 358–362, 1996.

[49] J. Gruber and P. Orszag. (2003, Mar.). What to do about the social security earnings test? Center for Retirement Research, *Issues in Brief ib-1*. [Online]. Available: http://ideas.repec.org/p/crr/issbrf/ib-1.html.

[50] J. Gruber and D. A. Wise, "Social security programs and retirement around the world," *Res. Labor Econ.*, vol. 18, pp. 1–40, 1999.

[51] A. L. Gustman and T. L. Steinmeier, "The social security retirement earnings test, retirement and benefit claiming," National Bureau of Economic Research, Cambridge, MA, Working Paper 10905, Nov. 2004.

[52] D. Guest, *The Emergence of Social Security in Canada*, 3rd ed. ed. Vancouver, BC, Canada: UBC Press, 1997.

[53] M. Feldstein, "The optimal level of social security benefits," National Bureau of Economic Research, Cambridge, MA, Working Paper 0970, Aug. 1986.

[54] L. Friedberg, "The labor supply effects of the social security earnings test," *Rev. Econ. Statist.*, vol. 82, no. 1, pp. 48–63, 2000.

[55] M. Feldstein, "The missing piece in policy analysis: Social security reform," *Amer. Econ. Rev.*, vol. 86, no. 2, pp. 1–14, 1996.

[56] M. S. Feldstein and J. B. Liebman, Eds., *The Distributional Aspects of Social Security and Social Security Reform*. Chicago, IL: Univ. of Chicago Press, 2002.

[57] S. J. Haider and G. Solon. (2000). "Nonrandom selection in the HRS social security earnings sample," *RAND—Labor and Population Program*, Working Papers. [Online]. Available: http://econpapers.repec.org/RePEc:fth:randlp:00-01

[58] P. Henman and M. Adler. (2003). "Information technology and the governance of social security," [Online]. Available: http://espace.library.uq.edu.au/view/UQ:166078

[59] A. Hicks, "Qualitative comparative analysis and analytical induction: The case of the emergence of the social security state," *Sociol. Methods Res.*, vol. 23, no. 1, pp. 86–113, Aug. 1994.

[60] M. Hill, *Social Security Policy in Britain*. Cheltenham, U.K.: Edward Elgar, 1990.

[61] H. Huang, S. Imrohoroglu, and T. J. Sargent, "Two computations to fund social security," *Macroecon. Dyn.*, vol. 1, no. 1, pp. 7–44, 1997.

[62] M. D. Hurd, J. P. Smith, and J. M. Zissimopoulos, "The effects of subjective survival on retirement and social security claiming," *J. Appl. Econometr.*, vol. 19, pp. 761–775, 2004.

[63] A. Imrohoroglu, S. Imrohoroglu, and D. H. Joines, "Computing models of social security," *QM&RBC Codes, Quantitat. Macroecon. Real Bus. Cycles*, 1998.

[64] J. Kim, "The impact of e-government on child support enforcement policy outcomes," in *Proc. 8th Annu. Int. Conf. Digital Govern. Res.*, 2007, pp. 212–221.

[65] W. van der Klaauwa and K. I. Wolpinb, "Social security and the retirement and savings behavior of low-income households," *J. Econometr.*, vol. 145, pp. 21–42, 2008.

[66] L. J. Kotlikoff, K. Smetters, and J. Walliser, "Distributional effects in a general equilibrium analysis of social security," in *The Distributional Aspects of Social Security and Social Security Reform* (ser. NBER Chapters). Cambridge, MA: Nat. Bureau Econ. Res., 2002, pp. 327–370.

[67] A. B. Krueger and J.-S. Pischke, "The effect of social security on labor supply: A cohort analysis of the notch generation," *J. Labor Econ.*, vol. 10, no. 4, pp. 412–437, 1992.

[68] H.-C. M. Kum, D. Duncan, K. Flair, and W. Wang, "Social welfare program administration and evaluation and policy analysis using knowledge discovery and data mining (KDD) on administrative data," in *Proc. Annu. Nat. Conf. Digital Govern. Res.*, 2003, pp. 1–6.

[69] H.-C. M. Kum, D. Duncan, and W. Wang, "Understanding social welfare service patterns using sequential analysis," in *Proc. Annu. Nat. Conf. Digital Govern. Res.*, 2004, pp. 1–2.

[70] C. Mesa-Lago, "Social security in Latin America," in *Latin Amer. Res. Rev.*, 2007.

[71] L. J. Kotlikoff, K. A. Smetters, and J. Walliser, "Social security: Privatization and progressivity," National Bureau of Economic Research, Cambridge, MA, Working Paper 6428 Feb. 1998.

[72] R. Lee and S. Tuljapurkar, "Stochastic forecasts for social security," in *Frontiers in the Economics of Aging* (ser. NBER Chapters). Cambridge, MA: Nat. Bureau Econ. Res., 1998, pp. 393–428.

[73] D. R. Leimer and S. D. Lesnoy, "Social security and private saving: New time-series evidence," *J. Political Econ.*, vol. 90, no. 3, pp. 606–629, 1982.

[74] D. R. Leimer, "Lifetime redistribution under the social security program: A literature synopsis," *Soc. Secur. Bull.*, vol. 62, no. 2, pp. 43–51, 1999.

[75] J. C. Leung, "Social security reforms in China: Issues and prospects," *Int. J. Soc. Welfare*, vol. 12, pp. 73–85, 2003.

[76] T. R. Marmor and J. L. Mashaw, "Understanding social insurance: Fairness, affordability, and the modernization of social security and medicare," *Health Affairs*, vol. 25, no. 3, pp. w114–w134, 2006.

[77] D. McAullay, G. Williams, J. Chen, H. Jin, H. He, R. Sparks, and C. Kelman, "A delivery framework for health data mining and analytics," in *Proc. 28th Australasian Conf. Comput. Sci.*, 2005, pp. 381–387.

[78] J. Millar, *Understanding Social Security: Issues for Policy and Practice*, 2nd ed. ed. Bristol, U.K.: Policy Press, 2009.

[79] O. S. Mitchell and J. W. Phillips, "Retirement responses to early social security benefit reductions," National Bureau of Economic Research, Cambridge, MA, Working Paper 7963, Oct. 2000.

[80] O. S. Mitchell and J. W. Phillips, "Social security replacement rates for alternative earnings benchmarks," Retirement Research Center, Univ. Michigan, Ann Arbor, Working Paper wp116, May 2006.

[81] J. A. Olson, "Linkages with data from social security administrative records in the health and retirement study," *Social Security Bulletin*, 1999.

[82] E. Ooghe, E. Schokkaert, and J. Flechet, "The incidence of social security contributions: An empirical analysis," *Empirica*, vol. 30, no. 2, pp. 81–106, 2003.

[83] L. G. Pee and A. Kankanhalli, "Understanding the drivers, enablers, and performance of knowledge management in public organizations," in *Proc. 2nd Int. Conf. Theory Practice Electron. Govern.*, 2008, pp. 439–466.

[84] J. F. Quinn and R. V. Burkhauser, "Influencing retirement behavior: A key issue for social security," *J. Policy Anal. Manag.*, vol. 3, no. 1, pp. 1–13, 1983.

[85] R. Rofman, "Social security coverage in Latin America," *Social Protection Series Discussion Paper*, May 2005.

[86] S. Rosen and P. Taubman, "Changes in life-cycle earnings: What do social security data show?," *J. Human Res.*, vol. 17, no. 3, pp. 321–338, 1982.

[87] N. Rossi and I. Visco, "National saving and social security in Italy," *Ricerche Economiche*, vol. 49, pp. 329–356, 1995.

[88] K. Rowlingson and Policy Studies Institute Staff, *Social Security Fraud: The Role of Penalties*. London, U.K.: Stationery Office Books, 1997.

[89] B. Rubenstein-Montano, J. Buchwalter, and J. Liebowitz, "Knowledge management: A U.S. social security administration case study," *Govern. Inf. Q.*, vol. 18, no. 3, pp. 223–253, 2001.

[90] J. Rust and C. Phelan, "How social security and medicare affect retirement behavior in a world of incomplete markets," EconWPA, Washington, DC, Working Paper Public Economics 9406005, Jun. 1994.

[91] A. A. Samwick, "New evidence on pensions, social security, and the timing of retirement," National Bureau of Economic Research, Cambridge, MA, Working Paper 6534, Apr. 1998.

[92] E. Sheshinski and Y. Weiss, "Uncertainty and optimal social security systems," *Quart. J. Econ.*, vol. 96, no. 2, pp. 189–206, May 1981.

[93] S. Rosen and P. Taubman, "Changes in life-cycle earnings: What do social security data show?," *J. Human Res.*, vol. 17, no. 3, pp. 321–338, 1982.

[94] J. M. Smith, "Viewpoint on public service and computer science," *Commun. ACM*, vol. 52, no. 11, pp. 34–35, 2009.

[95] J. E. Stiglitz, *Economics of the Public Sector*, 3rd ed. New York: Norton, 2000.

[96] A. Börsch-Supan and R. Schnabel, "Social security and declining labor-force participation in Germany," *Amer. Econ. Rev.*, vol. 88, no. 2, pp. 173–178, 1998.

[97] P. Orszag and J. E. Stiglitz, *Rethinking Pension Reform: Ten Myths About Social Security Systems*, World Bank, 1999.

[98] L. H. Thompson, "The social security reform debate," *J. Econ. Literature*, vol. 21, no. 4, pp. 1425–1467, 1983.

[99] C. Usher, J. Wildfire, and S. Schneider, "Family to family. Tools for rebuilding foster care: The need for self evaluation—Using data to guide policy and practice," Nat. Assoc. College Admission Counseling, Arlington, VA, Rep., 2001.

[100] C. L. Usher, E. Locklin, J. B. Wildfire, and C. C. Harris, "Child welfare performance ratings: One state's approach," *Administrat. Soc. Work*, vol. 25, pp. 35–51, May 2001.

[101] D. Webster, B. Needell, and J. Wildfire, "Data are your friends: Child welfare agency self-evaluation in Los Angeles county with the family to family initiative," *Children Youth Serv. Rev.*, vol. 24, nos. 6–7, pp. 471–484, 2002.

[102] B. C. Williams, L. B. Demitrack, and B. E. Fries, "The accuracy of the national death index when personal identifiers other than social security number are used.," *Amer. J. Public Health*, vol. 82, no. 8, pp. 1145–1147, Aug. 1992.

[103] S. Wu, Y. Zhao, H. Zhang, C. Zhang, L. Cao, and H. Bohlscheid, "Debt detection in social security by adaptive sequence classification," in *Proc. 3rd Int. Conf., Knowl. Sci., Eng. Manage.* (ser. Lecture Notes in Computer Science 5914). New York: Springer-Verlag, 2009, pp. 192–203.

[104] P. Zhang, F. Xu, L. Jiang, and R. Ge, "G2c e-government: Shanghai social security and citizen services," in *Proc. 7th Int. Conf. Electron. Commerce*, 2005, pp. 558–563.

[105] H. Zhang, Y. Zhao, L. Cao, and C. Zhang, "Class association rule mining with multiple imbalanced attributes," in *Australian Conf. Artif. Intell.*. New York: Springer-Verlag, 2007, pp. 827–831, (ser. Lecture Notes in Computer Science 4830).

[106] H. Zhang, Y. Zhao, L. Cao, and C. Zhang, "Combined association rule mining," in *Proc. 12th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining*, 2008, pp. 1069–1074.

[107] H. Zhang, Y. Zhao, L. Cao, and C. Zhang, "Rare class association rule mining with multiple imbalanced attributes," in *Proc. 20th Australian Joint Conf. Adv. Artif. Intell.*, 2009, pp. 827–831.

[108] H. Zhang, Y. Zhao, L. Cao, C. Zhang, and H. Bohlscheid, "Customer activity sequence classification for debt prevention in social security," *J. Comput. Sci. Technol.*, vol. 24, no. 6, pp. 1000–1009, 2009.

[109] J. Zhang and J. Zhang, "How does social security affect economic growth? evidence from cross-country data," *J. Populat. Econ.*, vol. 17, no. 3, pp. 473–500, 2004.

[110] Y. Zhao, H. Zhang, F. Figueiredo, L. Cao, and C. Zhang, "Mining for combined association rules on multiple datasets," in *Proc. Int. Workshop Domain Driven Data Mining*, 2007, pp. 18–23.

[111] Y. Zhao, H. Zhang, L. Cao, C. Zhang, and H. Bohlscheid, "Efficient mining of event-oriented negative sequential rules," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, 2008, pp. 336–342.

[112] Y. Zhao, H. Zhang, L. Cao, C. Zhang, and H. Bohlscheid, "Combined pattern mining: From learned rules to actionable knowledge," in *Advances in Artificial Intelligence (ser. Lecture Notes in Computer Science 5360)*. New York: Springer-Verlag, 2008, pp. 393–403.

[113] Y. Zhao, C. Zhang, and L. Cao, *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*. Hershey, PA: IGI Global, 2009.

[114] Y. Zhao, H. Zhang, L. Cao, H. Bohlscheid, Y. Ou, and C. Zhang, "Data mining applications in social security," in *Data Mining for Business Applications*, L. Cao, P. S. Yu, C. Zhang, and H. Zhang, Eds. New York: Springer, 2009, pp. 81–96.

[115] Y. Zhao, H. Zhang, S. Wu, J. Pei, L. Cao, C. Zhang, and H. Bohlscheid, "Debt detection in social security by sequence classification using both positive and negative patterns," in *Machine Learning and Knowledge Discovery in Databases* (ser. Lecture Notes Computer Science 5782). New York: Springer-Verlag, 2009, pp. 648–663.

[116] Y. Zhao, H. Zhang, L. Cao, C. Zhang, and H. Bohlscheid, "Mining both positive and negative impact-oriented sequential rules from transactional data," in *Advances in Knowledge Discovery and Data Mining* (ser. Lecture Notes Computer Science 5476). New York: Springer-Verlag, 2009, pp. 656–663.

[117] Z. Zheng, Y. Zhao, Z. Zuo, L. Cao, H. Zhang, and C. Zhang, "An efficient GA-based algorithm for mining negative sequential patterns," in *Proc. Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining*, 2010, pp. 262–273.

[118] Z. Zheng, Y. Zhao, Z. Zuo, and L. Cao, "Negative-GSP: An efficient method for mining negative sequential patterns," in *Proc. AusDM 2009*, pp. 63–67.

[119] Australian National Audit Office, *ANAO Centrelink Audit Report 2007–08*. Canberra, Australia: Commonwealth, 2008.

[120] Centrelink, *Integrated Activity Management Developer Guide*. Canberra, Australia: Commonwealth, 1999.

[121] Centrelink, *Annual Report 2009*. Canberra, Australia: Commonwealth, 2009.

[122] Department of Human Services, *Better Dealings with Government: Innovation in Payments and Information Services*. Canberra, Australia: Commonwealth, 2009.

[123] "The data measures, data composites, and national standards to be used in child and family services reviews, attachment b: Methodology for developing the composites," U.S. Department of Health and Human Services, Washington, DC, 2006.

[124] "Administration for children & families," child Welfare Monitoring. (2007, Nov.). [Online]. Available: http://www.acf.hhs.gov/programs/cb/cwmonitoring/index.htm

[125] (2007). *National Resource Center for Child Welfare Data and Technology* [Online]. Available: http://www.nrccwdt.org/index.html

[126] *Centrelink Customer Risk Rating at the Initial Registration*, UTS-Centrelink Contract Research Project, 2009.

[127] *The Provision of Income Reporting Data Analysis Services*, UTS-Centrelink Contract Research Project, 2006.

[128] *Pattern Analysis and Risk Control of E-Commerce Transactions to Secure Online Payments*, Australian Research Council Linkage Grant, 2007–2009.

[129] *Centrelink Fraud Investigation: Opportunities and Test*, UTS-Centrelink Contract Research Project, 2010.

[130] ARC Linkage, *Data Mining of Activity Transactions to Strengthen Debt Prevention*, Australian Research Council Linkage Grant, 2007–2009.

[131] ARC Linkage, *Detecting Significant Changes in Organisation Customer Interactions Leading to Non-Compliance*, Australian Research Council Linkage Grant, 2010–2013.

[132] *Detecting Incorrect Income Declaration in Real Time*, UTS-Centrelink Contract Research Project, 2010.

[133] [Online]. Available: http://www.centrelink.gov.au/internet/internet.nsf/about_us/fraud_index.htm

[134] (2009). [Online]. Available: http://www.skipease.com/blog/data-mining/data-mining-detects-welfare-fraud

Longbing Cao (SM'06) received the Ph.D. degrees in intelligent sciences from the Chinese Academy of Sciences, China, the Ph.D. degree in computing sciences from the University of Technology Sydney, Sydney, Australia.

Since 2004, he has been leading the development of social security and social welfare data mining. He is currently a Professor and the Director of Advanced Analytics Institute, University of Technology Sydney, Sydney, N.S.W., Australia. He is also the Data Mining Research Leader of the Australian Capital Markets Cooperative Research Centre. He has prodigious experience in practical innovation in enterprise data mining and analytics in many different domains, including the public sector, social welfare, capital markets, banking, insurance, telecommunication, and education. His research interests include data mining and machine learning and their applications, behavior informatics, multiagent technology, and agent mining.