Expert Systems with Applications 36 (2009) 9819–9833

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A proximate dynamics model for data mining

Yunfei Yin

School of Automation Science and Electrical Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100083, China

ARTICLE INFO

Keywords:
Data mining
Dynamics
Stability
Clustering
Proximate model

ABSTRACT

Many association rules with low supports and high confidence are commonly not convincing, so how to enhance the conviction of such rules is a big issue. In this paper, we explore the dynamics features of the domain dataset, and by predefining some dynamics parameters (as priori knowledge), we construct a proximate dynamics model for data mining so as to enhance the conviction of the rules with low supports. For constructing such dynamics model for data mining, we adopt three techniques: (1) a large domain dataset is classified into several sub-clusters, which we regard as proximate dynamics systems; (2) the data mining process is a solving process of differential equations, which captures the changes of the data, not only the values themselves; and (3) a weighting method is used to synthesize the local mining results with the users' preferences. Although we "arbitrarily" apply the dynamics parameters into the sub-clusters of the given dataset, the experimental results are very well by comparing with FP-growth algorithm and CLOSET+ algorithm. Experiments conducted on the distributed network with three real life datasets show that the proposed method can discover the knowledge based on dynamics, which is potentially useful for mining the rules with low supports and high confidence.

1. Introduction

1.1. Problem description and motivation

Taking supermarket shopping model as an example, conventional data mining can be used to analyze the relationship between purchased commodities. Unfortunately, it cannot discover valid business strategies on different groups of customers. Therefore, one important fundamental problem is how to classify the customers into different segments according to their spending behaviors.

In the paper, we propose a dynamics model for data mining, where we capture the spending behaviors of customers as the parameters of a single pendulum model. And further, simulating the movements of the single pendulum, we also consider the spending behaviors of customers to be several vibrations like a single pendulum. Although the relations between the spending behaviors of customers and the vibrations of the single pendulums are somewhat arbitrarily, the experimental results are satisfying, especially for the three selected datasets, viz., FoodMart dataset, stock experimental dataset and aircraft dataset. Finally, by such dynamics model, we can obtain some useful dynamics patterns about the spending behaviors of customers.

So, our motivation is to construct a dynamics model based on single pendulum for data mining, and the maximum consumption of the customer is mapped into the maximum amplitude of the single pendulum, the drop of the customer consumption is mapped into the damped vibration of the single pendulum, the first consumption of the customer is mapped into the initial state of the single pendulum, and others are by the same token. Finally, through such dynamics model, we investigate its dynamics properties and discover the useful dynamics patterns.

1.2. Related work

Related work about dynamics model of data mining includes three aspects: (1) applications of dynamics model, e.g. Mayaka, Stigter, Heitkonig, and Prins (2004) presented a dynamics model to compare six alternative management strategies with respect to their economic performance and impact on population size and structure of Buffon's kob (Kobus kob kob); (2) customer retention model, e.g. Larivière and Van den Poel (2005) presented three important measures of customer outcome: next buy, partial-defection and customers' profitability evolution, and by means of random forests techniques, a broad set of explanatory variables, including past customer behavior, observed customer heterogeneity and some typical variables were investigated, and finally, the findings suggested that past customer behavior was more important to generate repeat purchasing and favorable profitability evolutions, while the intermediary's role had a greater impact on the customers' defection proneness; and (3) data mining by priori knowledge, and there are two parts related to this aspect: one is mining useful rules with low supports (e.g. Koh, Rountree, & O'Keefe, 2008) and the other is post-mining (e.g. Zhang, Zhang, & Yan, 2003).

E-mail address: yinyunfei@asee.buaa.edu.cn

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.02.033

As to the first aspect of the related work: applications of dynamics model, there are many researches reported. An and Jeng provided heuristics and guidelines of developing system dynamics models based on given business process models along with associated reference contexts, and an supply chain management models to derive system dynamics models was also given (An et al., 2005). Luna-Reyes et al. described a dynamic theory of the socio-technical processes involved in the definition of an Integration Information problem in New York State (NYS) (Luna-Reyes, Andersen, Richardson, Pardo, & Cresswell, 2007). Panait and Tuyls provided a Replicator Dynamics model for traditional multiagent Q-learning, and extended the differential equations to account for lenient learners: agents that forgave possible mistakes of their teammates that resulted in lower rewards (Panait et al., 2007). Shuai et al. proposed a novel generalized particle dynamics model for software cybernetics in the context of optimal allocation of software resources and software jobs in complex environment, and the approach transformed software cybernetics problems into the kinematics and dynamics of particles in a force-field (Shuai, Zhang, & Lu, 2006). Shuai, Zhang, and Lu (2006) also presented a novel particle dynamics model for the adaptive self-organization of software processes in complex distributed environment, and in the model, the software processes in distributed environment was regarded as a random Markov process. Kagawa et al. considered C-posture as the equilibrium point attractor in the musculoskeletal dynamics of paraplegic patients (Kagawa, Fukuda, Fukumura, & Uno, 2005). Zhang et al. proposed a contact dynamics model of inkjet technology-based oligo DNA microarray spotting process, and the proposed dynamics model could reasonably well explain the dynamics of the oligo DNA microarray spotting process (Zhang, Ma, & Diao, 2006). Yang et al. utilized system dynamics theory to establish the dynamic model of demand side management (Yang, Zhang, & Tong, 2006).

To be followed, there are also many researches reported about customer retention model. Such as, Yin proposed a customer retention model based on bundled commodities mining, where the similarly fuzzy classification method was used to classify the customers (Yin, 2008). Chu et al. proposed a hybridized architecture to deal with customer retention problems and not only through predicting churn probability but also by proposing retention policies (Chu, Tsai, & Ho, 2007). Chellappa and Kumar argued that product competition on the Web was not for generic products but, rather, for expected and augmented product bundles, and found that even in the absence of price premiums, variance in the ability to offer online services could affect pricing strategies and possibly contribute to online price dispersion (Chellappa & Kumar, 2005). Ng and Liu proposed a solution that integrated various techniques of data mining, such as feature selection via induction, deviation analysis, and mining multiple concept-level association rules to form an intuitive and novel approach to gauging customer loyalty and predicting their likelihood of defection (Ng & Liu, 2000). Gupta and Kim applied a integrated structural equation modeling approach to decision support for customer retention in a virtual community, and the application results provided insights for practitioners on how to retain their customers (Gupta & Kim, 2008). Hidalgo et al. analyzed the convenience for the firm of improving customer retention by matching the lowest price in the Chilean private pension system, and the results suggested that matching the industry's price leader reduced the firm's customer lifetime value (Hidalgo, Manzur, Olavarrieta, & Farías, 2008). Sweeney and Swait argued that brand credibility on customer loyalty could have a significant role to play in managing long-term customer relationships (Sweeney & Swait, 2008). Saccani et al. discussed the configuration of the after-sales supply chain which was potential contribution to company profitability, customer retention and product development (Saccani, Johansson, & Perona, 2007). Tsai and Huang drew on marketing and consumer behavior literature to formulate a conceptual framework that considered community-based, customization-based, desire-based, and constraint-based drivers of online customer retention (Tsai & Huang, 2007).

Finally, as to the third aspect of the related work: data mining by priori knowledge, the related work is as follows. El-Hajj et al. proposed an approach that allowed the efficient mining of frequent itemsets patterns, while pushing simultaneously both monotone and anti-monotone constraints during and at different strategic stages of the mining process (El-Hajj, Zaiane, & Nalos, 2005). Elhadary et al. proposed an efficient robust combined clustering technique using neural networks for large image databases that required the user to provide a maximum number of clusters (Elhadary, Tolba, Elsharkawy, & Karam, 2007). Zhang and Zhang proposed a multiagent data warehousing and multiagent data mining approach for brain modeling, and an algorithm named Neighbor-Miner was proposed for the approach with agent similarity as a priori knowledge (Zhang & Zhang, 2004). Wang et al. presented a gene expression programming decision tree system which could construct a decision tree for classification about the distribution of data, and the method could solve n-class problem in a single run (Wang, Li, Han, & Lin, 2009). Taher and El-Ghazawi, presented a technique suitable for multitasking and for cases of single applications that can change the course of processing in a non-deterministic fashion based on data or priori knowledge (Taher et al., 2006). Qu, Wang, and Liu (2004) presented a model for data mining from spectra library by using the filed measured data to drive the model, and the obtained crop structure variables and its probability distribution were regarded as priori knowledge. Kuramochi and Karypis proposed a computationally efficient algorithm for finding frequent patterns corresponding to geometric subgraphs in a large collection of geometric graphs with low support values (Kuramochi & Karypis, 2007). Bodon and Rónyai presented a version of the TRIE data structure which outperformed hash-trees in some data mining applications, and also presented a simpler and scalable algorithm which turned out to be faster for lower support thresholds (Bodon & Rónyai, 2003).

1.3. Contributions

Since dynamics has been used in data mining as priori knowledge, we can also construct a customer retention model to maximize the potential benefits for the companies. Further, it is interesting to note that, data mining based on dynamics can form different movements such as lightly damped vibration, over damped vibration, critical damped vibration, and so on.

Therefore, the main contributions of the dynamics model for data mining are: (1) Mining "system" from dataset. Traditionally, data mining can discover rules, patterns, trends, and so on, but it cannot discover entire "system" from the dataset. In our work, by predefining dynamics system parameters, we can discover or optimize an entire "system" which consists of a series of parameters. (2) Target customer classification. For example, for the supermarket customers living nearby, their spending behaviors can be proximately modeled into the lightly damped vibration of a single pendulum; for the temporary customers of supermarket, their spending behaviors can be proximately modeled into the over damped vibration of a single pendulum; and for the hesitating customers of supermarket, the spending behaviors can be proximately modeled into the critical damped vibration of a single pendulum. (3) Stability analysis for dynamical data mining. For the complex dataset with changing scale, how to find valuable patterns and how to find stable laws are big issues. By constructing the dynamics model for data mining, we may find some solutions about these from the viewpoint of dynamics theory.