Expert Systems with Applications 40 (2013) 2137–2146

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

An enhanced Customer Relationship Management classification framework with Partial Focus Feature Reduction

Yan Tu, Zijiang Yang *

School of Information Technology, York University, Toronto, Ontario, Canada

ARTICLE INFO

Keywords:
Customer Relationship Management
Classification
Feature selection
Imbalanced classification
Ensemble classification

ABSTRACT

Effective data mining solutions have for long been anticipated in Customer Relationship Management (CRM) to accurately predict customer behavior, but from various industry research and case studies we have observed sub-optimal CRM classification models due to inferior data quality inherent to CRM data set. In this paper, one type of CRM data with a distinctive distribution pattern of Reduced Dimensionality is discussed. A new classification framework termed Partial Focus Feature Reduction is proposed to resolve CRM data set with Reduced Dimensionality using a collection of efficient data mining techniques characterizing a specially tailored modality grouping method to significantly improve data quality and feature relevancy after preprocessing, eventually achieving excellent classification performance with the right combination of classification algorithms.

1. Introduction

Customer Relationship Management is "the strategic use of information, processes, technology, and people to manage the customer's relationship with your company (Marketing, Sales, Services, and Support) across the whole customer life cycle" (Kincaid, 2003). The technological advancement has enabled new approaches – notably data mining – to be applied for finding the best CRM strategies. Ngai, Xiu, and Chau (2009) argue analytical CRM as a sub category of CRM, where data mining can play an essential role in analyzing customer data. However, there are factors imposed by industry nature that constitute major challenges for building high performance CRM classification models in the real-world application. Data quality is a salient issue for CRM classification practitioners in that various types of data anomaly largely complicate the data preparation and classification processes. Another complication arising from the cascading types of data anomaly is that no generally accepted data mining classification procedure can be established since it is hard to find one methodology that addresses all common data mining problems possessed by CRM data. Typically, a difficult CRM classification problem should involve data sets that possess the following data mining challenges:

- Heterogeneous data
- High feature dimension
- Severe data anomaly
- Imbalanced classification
- Data scrambling

Imbalanced classification is known to result in biased models that ignore instances belonging to the minority class. Although many methods are proposed to address this issue, the solutions are still case-dependent at the best (García, Sánchez, & Mollineda, 2012). Data collection processes in CRM such as survey and questionnaire are susceptible to missing data when customers are reluctant to provide some information or simply do not have chance to do so; human errors and misinterpretation also add up to noise in data. Moreover, customer information are heterogeneous with different measurement scale (e.g. the numeric feature of age and nominal feature of occupation), many of which are usually irrelevant to the target feature and will compromise classification performance if not filtered. With the ever increasing importance of CRM in every industry domain, CRM classification practitioners demand a standardized framework with streamlined data mining processes capable of delivering satisfactory result for general CRM data with all aforementioned data mining challenges.

Scholars have long devoted to the application of data mining in the CRM field. Hwang, Jung, and Suh (2004) propose a customer Life Time Value (LTV) model and customer segmentation framework; Chu, Tsai, and Ho (2007) propose a hybrid data mining model for the customer retention problem. However, results from researches targeted to general CRM classification problems are yet to be satisfactory. For example, the UC San Diego Data Mining Competition presented an imbalanced classification problem of target customer identification which ended up with the best score of only 68.5% AUC (UCSD, 2010). This paper proposes a new

* Corresponding author.
E-mail address: zyang@yorku.ca (Z. Yang).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.10.028

classification framework that targets the quality aspect of CRM data set with distinctive solution to a common but not thoroughly researched data anomaly of Reduced Dimensionality. The proposed methodology is tested on another real-world data set of KDD Cup 2009 Small Challenge and result compared with those from renowned data mining practitioners such as the IBM Research Lab (IBM Research, 2009) and University of Melbourne (Guyon et al., 2009). It is proved that the proposed methodology possesses superior performance to all competitors with generalized model that yields high accuracy with minimal training data.

The rest of the paper is organized as follows: Section 2 introduces detailed discussion of the proposed methodology. Section 3 presents the overview of classification data followed by classification result and discussion with respect to those from other researchers. Section 4 concludes the article and discusses future work to improve the proposed methodology.

2. Methodology

2.1. Reduced Dimensionality

There is one distinct characteristic common to some imbalanced CRM data sets that is significant: the instances of minority (usually the positive) class of the target feature often reside in a *Reduced Dimensionality* much smaller than of the majority (the negative) class in the problem's feature space, that is, the cardinality of all the features for the subset of positive instances are much smaller than for the negative instances. In formal definition, we have an imbalanced binary classification data set i with n instances: $\{(x_1,y_1)\ldots(x_n,y_n)\}$, where x represents the feature vector of one particular instance $D_i \in D$, and y the corresponding class label so that $y_i \in \{-1,+1\}$. If Y^+ represents the set of instances with positive class label, and Y^- the set of instances with negative class label, the data set D satisfies that $|Y^+| \ll |Y^-|$. Suppose A is the feature space of D, A^+ and A^- the subset feature space corresponding to the positive and negative class respectively. The ***Reduced Dimensionality*** is so defined that for $A_i \in A$, the cardinality of A_i^+ is far smaller than that of A_i^-:

$$\forall A_i, \quad |A_i^+| << |A_i^-|, \quad 1 \leqslant i \leqslant n$$

The degree of high overall dimensionality can be attributed to noises/outliers and the array of abundant subject attributes available in the problem domain, which are common to CRM data. However, A^+ appears often more condensed into a small dimension in i and is usually less affected by data anomalies. Moreover, it is costly to lose information representative of the minority class and somehow "desired" to overfit A^+ in a degree higher than the overall feature space so as to counter the disadvantage of data imbalance and to achieve better classification accuracy on the minority class.

2.2. Partial Focus Feature Reduction

The Reduced Dimensionality has hinted a possible preprocessing vehicle. In this proposed classification framework, a data mining workflow has been developed to exploit this phenomenon. Firstly, a new supervised binning method called the *Modality Grouping with Partial Focus* is introduced. This special binning method targets the nominal features in the data set rather than the numeric ones which creates bin for every value of a target nominal feature that has instances belonging to the minority class and then merges other values into a surrogate value category. For a nominal feature A_i with k value categories, a value category j either retains its value or is converted into the surrogate value category if the value it represents is absent in the reduced dimension:

$$A'_{ij} = \begin{cases} A_{ij}, A_{ij} \in A_i^+ \\ A_{is}, A_{ij} \notin A_i^+ | A_i^+| = 0 \end{cases} 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant k$$

Here A'_{ij} is the equivalent nominal feature in the new data set after the modality grouping, and A_{is} the surrogate value category for feature A_i. The same process is applied to the numeric features of the data set after feature discretization; missing values in all features are labeled as another surrogate value category. In this way, the modality grouping can efficiently smooth out most noise values and outliers relative to the target minority class.

It is challenging work to find the best feature selection method for a classification problem as it has to yield satisfactory performance and efficiency as well. For example, Bermejo et al. propose a variant of Wrapper Subset Selection (Bermejo et al., 2012) that addresses the problem of low efficiency which offset the superior performance of the original wrapper method (Kohavi & John, 1997). Here we use a feature reduction approach that takes a step further into converting the sub dimension of individual feature to retain only the most relevant sub-features with predictable efficiency. To do that, the resulting features from last step are binary-encoded into a feature pool and are put against an Information Gain ranking (Quinlan, 1986):

$$I(S,A) = E(S) - \sum_{j=1}^{k} \frac{|S_j|}{|S|} E(S_j)$$

Here I represents the Information Gain calculated upon splitting the data set S given feature A; E represents the entropy measure and $E(S_j)$ the entropy of subsets of S corresponding to j, one of A's k value categories. Only the collection of most relevant features is retained according to a cut-off threshold α:

$$F = \{f | I(f) \geqslant I(f)_{\max} * \alpha, \quad 0 < \alpha < 1\}$$

Here F represents the set of features to be retained, $I(f)$ the Information Gain of a certain feature f and $I(f)_{\max}$ the maximum Information Gain score in the feature pool.

At the end of the workflow, the classification framework employs the C4.5 Decision Tree (Quinlan, 1993) highly compatible with the preprocessed data to perform classification, and ensemble classifiers will be used for possible performance improvement. Although choice of classification algorithms should not be limited, C4.5 is found working best with this methodology, and the Bagging ensemble algorithm (Breiman, 1996) is chosen for the same reason. This framework is coined the *Partial Focus Feature Reduction* which demonstrates great capability resolving low quality imbalanced CRM data.

2.3. Methodology Implementation

The methodology implementation is primarily based on the popular WEKA data mining package (Hall et al., 2009). The WEKA Data Mining Package is open-source Java software package developed by the researchers of University of Waikato which offers a wide collection of data mining algorithms for classification and clustering as well as a variety of data mining utilities for data preprocessing and feature selection, etc. These data mining apparatuses are used for most of the work done in this paper. Particularly, the WEKA GUI is used for some initial preprocessing steps and the final classification of the preprocessed data sets. Analyses such as feature ranking are performed inside WEKA GUI and the results exported into external software such as Microsoft Excel for visualization if necessary.

Nonetheless, there are several operations in WEKA where memory management is sub-optimal. The first case of such is file I/O. WEKA uses a utility class called *ArffSaver* to perform simple writing of data set instances to file, which crashes the Java JVM when