Knowledge-Based Systems 91 (2016) 275–286

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Three-way recommender systems based on random forests

Heng-Ru Zhang, Fan Min*

School of Computer Science, Southwest Petroleum University, Chengdu 610500, China

ARTICLE INFO

Article history:
Received 12 December 2014
Revised 26 March 2015
Accepted 25 June 2015
Available online 27 June 2015

Keywords:
Cost sensitivity
Random forests
Recommender systems
Three-way decision

ABSTRACT

Recommender systems attempt to guide users in decisions related to choosing items based on inferences about their personal opinions. Most existing systems implicitly assume the underlying classification is binary, that is, a candidate item is either recommended or not. Here we propose an alternate framework that integrates three-way decision and random forests to build recommender systems. First, we consider both misclassification cost and teacher cost. The former is paid for wrong recommender behaviors, while the latter is paid to actively consult the user for his or her preferences. With these costs, a three-way decision model is built, and rational settings for positive and negative threshold values α^* and β^* are computed. We next construct a random forest to compute the probability P that a user will like an item. Finally, α^*, β^*, and P are used to determine the recommender's behavior. The performance of the recommender is evaluated on the basis of an average cost. Experimental results on the well-known MovieLens data set show that the (α^*, β^*)-pair determined by three-way decision is optimal not only on the training set, but also on the testing set.

1. Introduction

Recommender systems (RSs) have been extensively studied to present items such as movies [20,21,40], music [15] to consumers. There are two main approaches to implementing RSs, memory based and model based [8]. Memory-based methods [13,51] employ the entire user-item database to generate a prediction. Model-based methods [22,41,71,72] use demographic and content information to create a model that generates recommendations. Demographic RSs [14,48] generate recommendations based on compiled user demographic profiles, whereas content-based systems [5,25,35,49] treat recommending as a user-specific classification problem and learn a classifier for the user's likes or dislikes based on product features. Most existing RSs implicitly assume that the underlying task is one of binary classification, and so consider only two actions (recommend or not) for candidate items.

In this paper, we propose a framework that integrates three-way decision and random forests to build recommender systems. Three-way decision is introduced to modify the recommender's user output [4], that is, for a given item, to recommend, not recommend, or consult the user actively for his or her preferences. A misclassification cost is incurred for wrong recommender behaviors, including recommending movies to users who dislike them or not recommending movies to users who would like them. Teacher cost is incurred for the consultation, e.g., distributing coupons to users. Because two types of costs are considered, this approach essentially involves cost-sensitive learning [17,36,37,39,78]. There are various types of costs that can be considered [57,73]. In existing work, misclassification cost is the most widely addressed because classification is a major task of data mining (see, e.g., [28,60,77]). Test cost is often considered because data are not free (see, e.g., [58,61,69]). Delay cost is a major issue in the context of decision-theoretic rough sets [30,32,65,75]. Teacher cost is an essential consideration in active learning [6,43,54,55]. It is similar to test cost in that both "buy" data from the user. The major difference between the two lies in that test cost is incurred to obtain attribute values for data, whereas teacher cost is accrues to obtaining actual user decisions.

Three-way decision [33,34,67] is a methodological extension of decision-theoretic rough sets [31,59,62,64] that deals with situations where there are three possible decisions, namely, accept, reject, and wait and see [63,73,76]. It has been widely applied to situations such as filtering of spam email [76], determining biological matrices [18] and medical decision making [42]. The starting point is often a cost matrix with misclassification and delay costs. In this work, we consider teacher cost instead of delay cost. The purpose is also to compute two thresholds α^* and β^*: if the probability that a user will like a movie is greater than α^*, the movie is recommended., and if it is less than β^*, the movie is not recommended. Otherwise, we solicit his or her inclination, at the price of paying teacher cost.

To determine the recommender's behavior, we need to formulate the probability of the user's liking a movie. Decision-tree classification algorithms [50] are a natural approach. A decision tree

* Corresponding author. Tel.: +86 135 4068 5200.
E-mail address: minfanphd@163.com (F. Min).

http://dx.doi.org/10.1016/j.knosys.2015.06.019
0950-7051/

describes graphically the decisions to be made, the events that may occur, and the outcomes associated with combinations of decisions and events [11]. However, a single decision tree can only take advantage of limited information of users and items, and hence no suggestion is generated for many new users and new items. To overcome this limitation, we employ a random forest instead for this purpose. A random forest is a collection of decision trees [10], and different trees in the forest may include different information, therefore the missing of prediction seldom happens.

Our approach is depicted in Fig. 1. It is divided into three parts: (1) An optimal threshold pair (α^*, β^*) is computed based on the cost matrix, where α^* determines the probability necessary to recommend an item and β^* determines that necessary to not recommend an item. (2) The user, item, and rating information are merged to construct an aggregate decision table in which the rating information serves as the decision attribute. A random forest is then built from the aggregate training set, where each leaf is assigned a distribution of discrete ratings. The probability of each testing object is computed from the random forest. (3) Finally the RS selects an action based on P, α^*, and β^*.

Experiments on the well-known MovieLens data set (http://www.movielens.org/) show that (1) the threshold pair (α^*, β^*) determined with three-way decision is optimal not only on the training set, but also on the testing set and (2) our three-way RS outperforms the Pawlak, variable-precision, and probabilistic two-way models in terms of average cost.

The rest of the paper is organized as follows: Section 2 describes the three-way recommender problem and presents some background, including the rating system and cost-sensitive learning. An aggregate decision system is then built to mine the behavior of users on items. Section 3 discusses how to apply four rough-set models to our RSs. Section 4 constructs a random forest to predict the probability P and compute the average cost according to the three-way decision model. Section 5 presents experimental results from the MovieLens data set with the four rough-set models. How to set the threshold pair (α, β) is discussed in detail. Concluding remarks are presented in Section 6.

2. Problem statement

Most existing RSs take a rating system as input, and the recommender's accuracy is regarded as a kind of evaluation metric. Our three-way RS considers misclassification and teacher costs, and through cost-sensitive learning, we build proper classifiers to find a minimum average cost. For this purpose, the rating system is transformed into a decision system.

2.1. Rating system

We first revisit the definitions of information and rating systems [1,71].

Definition 1. An information system is a 2-tuple [71]

$$S = (U, A), \tag{1}$$

where $U = \{x_1, x_2, \ldots, x_n\}$ is the set of all objects, $A = \{a_1, a_2, \ldots, a_m\}$ is the set of all attributes, and $a_j(x_i)$ is the value of x_i with respect to attribute a_j for $i \in [1, n]$ and $j \in [1, m]$.

Example 2. Fig. 2 (a) shows an information system where $U = \{u_1, u_2, u_3, u_4, u_5\}$ and $A = \{Age, Gender, Occupation\}$. UID is the user identification, which is usually not viewed as an attribute. Another example of an information system is given by Fig. 2(b).

Definition 3. Let $U = \{u_1, u_2, \ldots, u_n\}$ be the set of users of a RS and $V = \{m_1, m_2, \ldots, m_l\}$ be the set of all possible items that can be recommended to users. Then the rating function is defined as

(b) Item table

MID	Year	Comedy	Action	Drama
m_1	90s	0	0	1
m_2	90s	0	0	1
m_3	60s	1	0	0
m_4	70s,80s	1	1	1
m_5	60s	1	0	0

(a) User table

UID	Age	Gender	Occupation
u_1	A25	F	technican
u_2	A50	M	writer
u_3	A18	F	student
u_4	A25	F	technican
u_5	A25	F	technican

(c) Rating

	m_1	m_2	m_3	m_4	m_5
u_1	Y	N	0	Y	Y
u_2	N	Y	N	Y	0
u_3	N	N	0	N	N
u_4	Y	Y	N	Y	0
u_5	Y	N	0	N	Y

Fig. 2. Rating system.

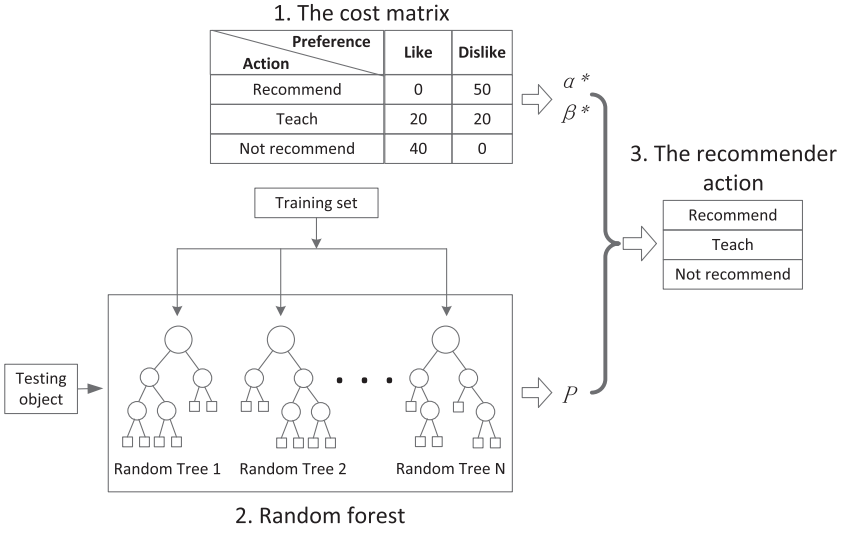

Fig. 1. Framework integrating three-way decision and a random forest.