Discovering Interesting Patterns Through User's Interactive Feedback

Presenter: Wei Yang

Outline

- ✓ Motivation
- ✓ Introduction
- ✓ Problem Statement
- ✓ Methodologies
- ✓ Experimental Study
- ✓ Conclusion

Motivation

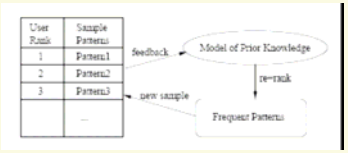
- ✓ Many patterns in the output while only a few of them is really interesting to a user.
- ✓ The measure of interestingness is subjective. There is no consistent objective measure to represent user's interest.

Introduction

- ✓ This paper introduces a new problem setting where the mining system interacts with the user, and proposes a framework to learn user's prior knowledge from interactive feedback. It also provides two models to represent a user's prior, and presents a two-stage approach to select sample patterns.
- ✓ Experiment results demonstrate the effectiveness of the approach and show that both models are able to learn user's background knowledge.

Introduction

- The system takes a set of candidate patterns as input.
- A model is created to represent a user's prior knowledge.
- At each round, a small collection of sample patterns are selected.
- The user ranks the sample patterns, and the feedback information is used to refine the model parameters.
- The system re-ranks the patterns according to the intermediate result and decide which patterns to be selected for next feedback.
- Finally, the top-ranked patterns are output as interesting patterns.



Problem Statement

- The interestingness of pattern P is determined by the difference between the observed frequency $f_0(P)$ and the expected frequency $f_e(P)$.
- Model the interestingness measure using two components: a model of prior knowledge and a ranking function.
- The model of prior knowledge M is used to compute the expected frequency of P as follows: $f_e(P) = M(P, \theta)$.
- A user feedback is formulated as a constraint on the model to be learned.
- The ranking function R is of the form: $R(f_0(P), f_e(P)) = \log f_0(P) - \log f_e(P)$, which returns the degree of interestingness of the pattern according to the observed frequency and the expected frequency.

Modeling Prior Knowledge

- Log – linear Model:
- Biased Belief Model

Log – linear Model

- Log – linear model is designed for item-set patterns.
- The log-linear model is used to study the frequency of an item-set comprising n-items: $f(x_1, x_2, \ldots, x_n)$.
- Given an item-set pattern $P = (i_1, \ldots, i_s)$, its expected frequency by a fully independent log-linear model is:

$$\log f_e(P) = u + \Sigma_{j=1,\ldots,s} u_j$$

Biased Belief Model

- The expected frequency of a pattern is determined by user's belief in the underlining data.
- Assign a belief probability to each transaction.
- A higher probability means the user is more familiar with this transaction. A lower one indicates that this transaction is novel to the user.
- The user's prior knowledge can be represented by a vector $[p_1,\ldots, p_m]$, where p_k is the belief probability for transaction k, and m is the total number of transactions.
- Given a pattern P, the value of $f_e(P)$ is proportional to the expected number of occurrences of P: $\Sigma_{k=1,\ldots, m}\ p_k * x_k(P)$, where $x_k(P) = 1$ if transaction k contains pattern P, otherwise, it is 0.

Sample Patterns Selection

Two stage approach:

- Progressive shrinking
- Clustering

Progressive Shrinking

- Define a shrinking ratio α ($0 < \alpha < 1$).
- At the beginning, the candidate set size N is equal to the size of the complete pattern collection.
- It gradually decreases to focus more on the highly ranked patterns.
- At each iteration, we update $N = \alpha N$, and the pattern set of clustering is the top-N patterns.

Clustering

- Suppose a user agrees to examine k patterns at each iteration, we cluster these top-N patterns into k clusters.
- Use Jaccard distance for clustering: given a pattern P1 and P2, the distance between P1 and P2 is defined as

$$D(P1, P2) = 1 - |T(P1) \cap T(P2)| / |T(P1) \cup T(p2)|$$

Where T(P) is the set of transactions which contain pattern P.

- The algorithm first picks an arbitrary pattern. While the number of picked patterns is less than k, the algorithm continues to pick a pattern which has the maximal distance to the nearest picked patterns.

Experimental Study

A series of experiments to examine the ranking accuracy.

- ✓ Item-set Patterns
- ✓ Sequential Patterns
- ✓ Sample Patterns Selection

Experimental Study – Item-set Patterns

- ✓ Run on a real data set *pumsb*.
- ✓ The accuracy of top 10% result of the log-linear model and biased belief model with different feedback size (5 or 10) is shown in Figure 2.
- ✓ Both models achieve higher than 80% (70%) accuracy with feedback size 10 (5).

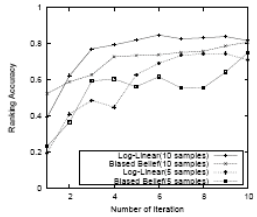


Figure 2: Top-10% ranking accuracy: k iterations.

Experimental Study – Sequential Patterns

- ✓ The accuracy of the top k percent (k=1,…, 10) ranking after 10 iterations is shown in Figure 3.
- ✓ The biased belief model works better than the log-linear model.
- ✓ The biased belief model gets 80% for top 10 percent rankings with fully ordered feedback.

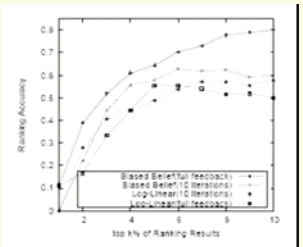


Figure 3: Top-k% ranking accuracy: 10 iterations.

Experimental Study – Sample Patterns Selection

- ✓ Compare strategies to select sample patterns for feedback.
- ✓ Selective sampling approach is comparatively worse.
- ✓ Top-N clustering approach is worse than shrinking and clustering method until the 5-th iteration.
- ✓ Shrinking and clustering approach is more efficient.

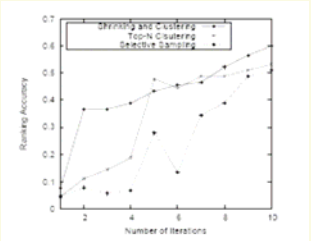


Figure 4: Top-10% ranking accuracy: sample pattern selection.

Conclusion

- ✓ This paper introduces a framework to learn user's prior knowledge from interactive feedback.
- ✓ Two models are proposed to represent a user's prior: the *log-linear model* and *biased belief model*.
- ✓ Finally, a two-stage approach is provided to select sample patterns for feedback: *progressive shrinking* and *clustering*.

Thank you!