A STUDY OF PROFIT MINING

Yu-Lung Hsieh and Don-Lin Yang
Dept. of Information Engineering and Computer Science, Feng Chia University, No. 100 Wenhwa Road, Taichung, Taiwan

Keywords: Financial data mining, Association rule, Inter-transaction, Profit mining, Risk, Win rate, Trading simulation.

Abstract: In the past decade, association rule mining has been used extensively to discover interesting rules from large databases. However, most of the produced results do not satisfy investors in the financial market. The reason for this is because association rule mining simply uses confidence and support to select interesting patterns while the investor is more interested in the result- trading at high profit and low risk. We propose a novel approach called Profit Mining which provides investors with trading rules including information about profit, risk, and win rate. To show the feasibility and usefulness of our proposal, we use a simple trading model of an inter-day trading simulation. This mining approach works well not only in the stock market, but also in the futures and other markets.

1 INTRODUCTION

In the last decade, many data mining researchers focused on the study of *association rule* (AR) mining (Agrawal, 1993; Shen, 1999). Especially in the financial applications, data mining can be used to make forecast (Boetticher, 2006). Using the application of AR on the stock market as an example, one can benefit from the rule: when the price of ABC stock goes up, the price of XYZ stock goes up with 70% confidence at the same day. However, this rule cannot depict the information of when to buy or when to sell the stock. As a result, the inter-transaction mining (ITM) was proposed (Lu, 1998) to improve the usage of AR. The generated rule from ITM indicates when the price of ABC stock goes up, then the price of XYZ stock goes up the next day with 70% confidence.

With ITM we still cannot be sure to have a profit as illustrated in the next example. Assume an investor earns 1 dollar from trading one unit of XYZ stock when both the prices of ABC and XYZ stocks go up. If the price of XYZ stock goes down today after the price of ABC stock went up yesterday, the loss is 3 dollars per unit where tax and trading fee are taken into account. Since the confidence of the generated rule is 70%, we can calculate the total loss of 2 dollars from ten trades. The calculation is 7 * 1 dollars – 3 * 3 dollars = -2 dollars. Although ITM provides the information of trading time, the chance of making a profit from simply following the generated rules is still small. The reason is that a high confidence rule does not necessary possess the benefit of high profit.

In the investment market, although every trader wants to make a profit, taking a loss is still inevitable and such a loss is called risk. If two trading approaches have the same profit and risk with two different winning chances of 1% and 99%, the investor would prefer the latter over the former. Here the winning chance is called WinRate.

To satisfy the above expectation of investors in the financial market, we propose a novel concept called *profit mining* (PM). The goal of PM is to mine the trading rules for the financial investors. Therefore, PM needs a trading model which can present various trading strategies used in the trading rules. The trading results from applying these rules are acquired by the trading simulation using the trading model, trading rules, and databases of historical transactions.

There are various trading rules and trading results of PM depending on the trading model and investor's expectations. In fact, there are a lot of trading models existed in PM and we only propose a simple model in Section 3.

The concept of PM is similar to utility mining (Yao, 2006) which uses a sales manager's perspective to reveal the profit that is important to the miner. Traditional statistical correlations may not measure how useful an itemset is in accordance with a user's preferences. Therefore, PM is a new mining approach like utility mining.

Hsieh Y. and Yang D..
A STUDY OF PROFIT MINING.
DOI: 10.5220/0003691705020506
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 502-506
ISBN: 978-989-8425-79-9

2 RELATED WORK

Agrawal (1993) proposed the association rule mining for finding the rule of related itemsets with high frequency and confidence from transactional database. Lu (1998) presented the inter-transaction association rules for mining the timing of stock prices going up or down. Zhang (2004) revealed "the fact that data mining in finance is involved with applications, data, and domain models leads to a conceptual framework consisting of three-dimensions." Boetticher (2006) performed several studies to mine the information from financial data in which the result is presented by profit, but lacks risk information. Risk management in financial market (Magdon-Ismail, 2004) is critical for the investors. In the real world applications, many investment experts use Trade Station (TradeStation, 2011) to build program models that can perform trading simulation. After the simulation, Trade Station generates reports to present the profit, risk and other trading results of the program model.

3 OUR PROPOSED PROFIT MINING

Profit mining consists of a trading model with the transactional database (TDB). According to the trading model and transactional database, the trading rules and trading results are generated. Using trading rules we can apply trading model to simulate the trading for the TDB and generate the trading result. First, we define a simple trading model called inter-day model as described below.

3.1 Trading Model – Inter-Day Model

Let MP = {None, Long, Short} be a set of market positions and TC = {Buy, Sell} be a set of trading commands. Let TO = {tc, qty, price} be a form of trading order where $tc \in TC$, $qty \in \{1,2\}$, $price \in R$ and $price>0$. For simplification, we limit the values of qty to 1 and 2. We say that TO is a BuyOrder (BO) where tc = "Buy", or a SellOrder (SO) where tc = "Sell". Let POS = {mp, hqty, hprice} be a form of hold position where $mp \in MP$, $hqty \in \{0,1\}$, $hprice \in R$ and $hprice \geq 0$. We say that POS is in a close position when mp = "None". Similarly, POS is in a long position when mp = "Long" or a short position when mp = "Short".

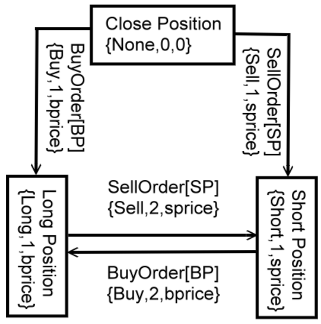

Figure 1: The state machine of our trading model.

Figure 1 shows these three hold positions in our model. A close position means that the investor does not hold any stock. "Long or Short Position" means that the investor expects the price of stock in the future to go up or down. BP and SP stand for Buy Pattern and Sell Pattern respectively, where *sprice* is the selling price and *bprice* is the buying price.

In the initial trading, an investor holds a close position. When a BP or SP occurs, the BO or SO is generated respectively. These orders change the close position to long or short position as indicated by the arrowhead lines from the top of Figure 10 to its left or right respectively.

If a long position POS_{t-1} (generated by TO_{t-1}) meets a BP, it is ignored until an SP is met to generate an SO. Then the long position POS_{t-1} is changed to the short position POS_t, where the arrowhead line is drawn from the long position (left) to the short position (right) in Figure 1. The TO_{t-1} is called a complete trade (CT) in which the POS_{t-1} is changed to the short position by TO_t. The case where the short position is changed to the long position is opposite to the last case.

Table 1: Transactional database (TDB).

TID	ItemSet	Price
1	<a,b>	500
2	<>	480
3	<a,b>	480
4	<b>	490
5	<a>	450
6	<>	410
7	<a,b>	440
8	<>	470

3.2 Transactional Database

Table 1 shows an example of transactional database with three attributes *TID*, *ItemSet* and *Price*. *TID* is a transaction ID possessing the time feature, where the item set (or event set) and trading price recorded at that time are *ItemSet* and *Price* respectively.

3.3 Trading Rule

When BP or SP occurs at the close position, there is a semantic ambiguity because the investor does not know which trading action to take, Buy or Sell? A new attribute of Trading Priority (TP) is required to solve the problem. Then, we have the format of the trading rule: <TP, BP, SP> where TP$\in$ {BF, SF}. BF is BuyFirst and SF is SellFirst.

To identify a pattern, we need to specify a *maxspan* value indicating the maximal number of transactions in the TDB containing the pattern. For example, If *maxspan* = 2, there are 2 patterns {{*a*(0)},{*a*(0)*b*(-1)}} at transaction TID 5. The pattern {*a*(0)*b*(-1)} has 2 items (or events) *a* and *b*. The number in the parenthesis of *b*(-1) is the interval which describes the distance from the base transaction TID 5 to the item *b* at transaction TID 4, which has a distance value of -1. A negative value means the backward distance. Since the pattern {*a*(0)*b*(-1)} occurs at TID 5, the trading price 450 of the pattern {*a*(0)*b*(-1)} is set at transaction TID 5, not at transaction TID 4.

3.4 Trading Results of Profit, Risk and WinRate

In our PM model one must specify the handling *fee* for each trading to cover the fee and tax as in the real world.

3.4.1 Profit

Let DB_{Begin} and DB_{End} be two TIDs which represent the TID of the first and last transactions in the TDB respectively. Let P be a function for getting the price at a specific TID, denoted as P(TID). Let T_i and T_j be two TIDs, where $T_i<T_j$ and $T_i,T_j \in$ TID. We denote that *i* and *j* are two trading orders and T_i and T_j are the locations of trading orders *i* and *j*, respectively. Let $minP(T_i, T_j)$ and $maxP(T_i, T_j)$ be two functions for getting the minimal and maximal prices from T_i to T_j, where $T_i<T_j$ and $T_i,T_j \in$ TID. Then the equation of NetProfit (NP) and *profit* of trading rule are as follows:

$$NP_{i,j} \begin{cases} P(T_j) - P(T_i) - 2*fee, \text{ if } mp_i = \text{“Long”} \\ P(T_i) - P(T_j) - 2*fee, \text{ if } mp_i = \text{“Short”} \end{cases} \quad (1)$$

$$Profit = \Sigma NP_{i,j} \quad (2)$$

3.4.2 Risk

To define the risk of trading rule, three variables are required: Consecutive Loss (CLoss), Draw Down (DD) and Run Up (RU). Their initial and maximal values are all 0. CLoss records the consecutive loss of net-profit by using the following equation:

$$\text{CLoss at time t: } CLoss_t = CLoss_{(t-1)} + NP_t \quad (3)$$

where $CLoss_0 = 0$ and $CLoss_t = 0$ if $(CLoss_{(t-1)} + NP_t) > 0$

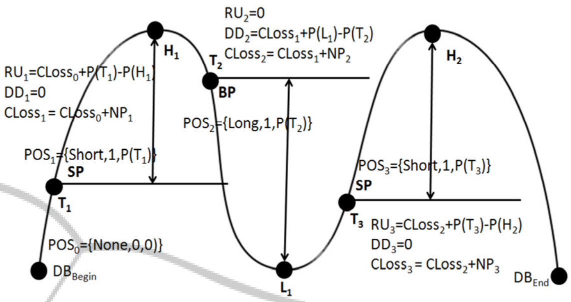

Figure 2: Computing the risk of trading rule using stock price vs. trading time curve.

We use Figure 2 to present the calculation of variables DD and RU which are used to record the risk during the trading process as MP = “Long” or MP = “Short”. DD records the difference between the buying price and the lowest price. One example is shown in the T_2 point of Figure 2. The L_1 point is the lowest point between T_2 and T_3. The equation to compute DD is shown below:

$$DD_{t-1} = CLoss_{t-2} + minP(T_{t-1}, T_t) - P(T_{t-1}) \quad (4)$$

$$\text{if } MP_{t-1} = \text{“Long” and } t>0 \text{ and } T_t \neq \text{Null}$$

However, if the BP at T_t is next to the DB_{End}, there is no SP to expect and equation (5) is used to compute DD. Otherwise, the value of DD is 0.

$$DD_t = CLoss_{t-1} + minP(T_t, DB_{End}) - P(T_t) \quad (5)$$

$$\text{if } MP_{(t-1)} = \text{“Long” and } t>0 \text{ and } T_t \neq DB_{End}$$

The definition of RU is similar to DD, but RU works at MP = “Short” and records the difference between the selling price and the highest price. The following two equations (6) and (7) corresponding to equations (4) and (5) are used for RU respectively.

$$RU_{t-1} = CLoss_{t-2} + P(T_t\text{-}1) - maxP(T_{t-1}, T_t) \quad (6)$$

$$\text{if } MP_{t-1} = \text{“Short” and } t > 0 \text{ and } T_t \neq \text{Null}$$

$$RU_t = CLoss_{t-1} + P(T_t) - maxP(T_t, DB_{End}) \quad (7)$$

$$\text{if } MP_{t-1} = \text{“Short” and } t > 0 \text{ and } T_t \neq DB_{End}$$

The current risk is the maximal absolute value among the current values of CLoss, DD, RU and the previous risk. The equation of risk is defined as follows:

Risk at time *t*: $Risk_t = \max[\ |CLoss_t|, |DD_t|, |RU_t|, Risk_{(t-1)}\ ]$ if $t > 0$ and $Risk_0 = 0$ if $t = 0$ (8)

3.4.3 WinRate

The variable WinRate is the ratio between the number of complete trades with net-profit >0 and the total number of complete trades (CTs).

The WinRate of trading result is defined below:

$$(\text{Total \# of CTs with NP>0}) / (\text{Total \# of CTs}) \times 100\% \quad (9)$$

3.5 Profit Rule

Let *minProfit*, *maxRisk* and *minWinRate* be user specified threshold values. We define a profit rule to be a trading rule *R* with trading results $\{Profit_R, Risk_R, WinRate_R\}$ if $minProfit \geq Profit_R$, $maxRisk < Risk_R$ and $minWinRate \geq WinRate_R$.

4 AN EXAMPLE OF MINING PROFIT RULES

We use a trading rule <BF, a(0), b(0)> from Table 1 to explain the trading simulation. As before, we set the trading fee to 1. The trading process is shown in Table 2, where there are fifteen attributes as defined previously.

At the beginning, the position = {None, 0, 0}. The first trading order is BuyOrder = {Buy, 1, 500}, because BP and SP occur at TID 1. Since the semantic ambiguity of trading appears, TP = BuyFirst is adopted. After trading, the position is changed from {None, 0, 0} to {Long, 1, 500} as shown in the Record No. 1 of Table 2. In the second trade at Record No. 2, BP and SP patterns occur at TID 3, where BP is ignored and the sell-order = {Sell, 2, 480} changes the position to {Short, 1, 480}. The NP_1 at Record No. 1 equals to 480 – 500 – 2 * 1 = -22. The value of $CLoss_1 = CLoss_0 + NP_1 = 0 + (-22) = -22$. The value of DD_1 equals to minP(1, 3) – P(1) = 480 - 500 = -20. The value of $RU_1 = 0$ because MP_1 = "Long". The value of $Risk_1 = \max(|-22|, |-20|, |0|, 0) = 22$. The value of $WinRate_1$ is 0 because $NP_1 \leq 0$.

The next trading order is BuyOrder = {Buy, 2, 450} is at TID 5 and NP_2 equals to 480 – 450 – 2 * 1 = 28 at Record No. 3. Here $CLoss_2 = 0$ and $DD_2 = 0$ because $CLoss_1 + NP_2 = -22 + 28 = 6 > 0$ and MP_1 = "Short" respectively. Then we have $RU_2 = -22 + P(3) - maxP(3, 5) = -22 + 480 - 490 = -32$ and $Risk_2 = \max(|0|, |0|, |-32|, 22) = 32$. The $WinRate_2$ is 50% because $NP_2 > 0$. The last trading order is SellOrder = {Sell, 2, 440} is at TID 7. After selling two stocks, NP_3 = 440 – 450 -2*1 = -12. We have $CLoss_3 = 0 + (-12) = -12$ and DD_3 = 0 + minP(5, 7) – P(5) = 410 – 450 = -40. Then $RU_3 = 0$ because MP_2 = "Long". The $WinRate_3$ is 33% because $NP_3 \leq 0$.

There is no buy pattern BP coming after TID 7. However, it is not risk free because there are transactions from TID 7 to DB_{End} (i.e., TID 8). We have RU_4 = -12 + P(7) - maxP(7, 8) = -12 + 440 – 470 = -42 and Risk4 = max(|-12|, |0|, |-42|, 40) = 42. The Profit of the trading rule is the summation of NPs = (-22) + 28 + (-12) = -6. Therefore, the trading result for the trading rule <BF, a(0), b(0)> is {-6, 42, 33%}.

Table 2: The simulated trading of the rule <BF, a(0), b(0)>.

No	Trading				MP	HQTY	NP	CLoss	DD	RU	Trading Result		
	CMD	QTY	TID	Price							Profit	Risk	WinRate
0					None	0	0	0	0	0	0	0	0%
1	Buy	1	1	500	Long	1	-22	-22	-20	0	-22	22	0%
2	Sell	2	3	480	Short	1	28	0	0	-32	6	32	50%
3	Buy	2	5	450	Long	1	-12	-12	-40	0	-6	40	33%
4	Sell	2	7	440	Short	1			0	-42		42	

5 EXPERIMENTAL RESULT

To verify the correctness of our profit mining and validate the mining results to satisfy the investor's expectation in the financial market, we use the following experiment to show the trading rule and trading result.

5.1 Our Experiment

In the experiment, we assume that the trading fee is 1, the value of maxspan is 2, and the thresholds of

Table 3: The Mining Results of the experiment.

No	Trading Priority	Buy Pattern	Sell Pattern	Profit	Risk	WinRate
1	SF	a(0)b(0)	b(0)	74	0	100%
2	BF	a(0)b(0)	b(0)a(-1)	36	20	50%
3	SF	a(0)b(0)	b(0)a(-1)	36	20	50%
4	BF	a(0)b(0)	b(0)a(-1)b(-1)	36	20	50%
5	SF	a(0)b(0)	b(0)a(-1)b(-1)	36	20	50%
6	BF	a(0)b(0)	b(0)b(-1)	36	20	50%
7	SF	a(0)b(0)	b(0)b(-1)	36	20	50%
8	SF	b(0)	b(0)	74	0	100%
9	BF	b(0)	b(0)a(-1)	36	20	50%
10	SF	b(0)	b(0)a(-1)	36	20	50%
11	BF	b(0)	b(0)a(-1)b(-1)	36	20	50%
12	SF	b(0)	b(0)a(-1)b(-1)	36	20	50%
13	BF	b(0)	b(0)b(-1)	36	20	50%
14	SF	b(0)	b(0)b(-1)	36	20	50%

minProfit, *maxRisk* and *minWinRate* are 10, 30 and 30%, respectively. After mining the TDB of Table 1, we found fourteen profit rules as shown in Table 3. For Rule No. 8, the BP is equal to its SP meaning that investors must sell their stocks first when they hold no stock, or they will lose money. For the rules No. 2 and No. 3, their BP, SP and trading results are the same. It means that investors can adopt buy-first or sell-first since their trading results are the same either way.

5.2 Discussion

As a preliminary study on financial data mining, we use a simple trading model of inter-day model for trading simulation. With different trading models, one can derive many different trading rules under specific transaction databases.

In data mining research, one expects to mine the knowledge of any kind that users might be interested in. These useful results of knowledge can be discovered and presented in the forms of rules, patterns, or any other forms to meet users' expectation.

Similarly in profit mining, we think more types of profit rules can be defined and discovered in any of the financial sectors. Since investors have their own measurement criteria for mining the preferred trading rules, there must be more useful models of financial data mining to be investigated.

6 CONCLUSIONS

In this paper, we present a new data mining in the financial market, called profit mining, with initial results showing the feasibility and usefulness of our proposed model. There still remains much research to be investigated other than association rule mining and inter-transaction mining. Currently we are working on efficient algorithms for profit mining in various trading models. Our future research also includes solving the challenge of reducing the search space and speeding up the mining process with limited memory space.

ACKNOWLEDGEMENTS

This research was supported by the National Science Council, Taiwan, under grant NSC 98-2221-E-035-059-MY2.

REFERENCES

Agrawal, R., Imielinski, T., and Swami, A. 1993. Mining association rules between sets of items in large databases, In *Proceedings of ACM SIGMOD*, pp. 207–216.

Boetticher, G.D., 2006. Teaching Financial Data Mining using Stocks and Futures Contracts, *Journal of Systemics, Cybernetics and Informatics,* Vol. 3, No. 3, pp. 26-32.

Lu, H., Han, J. and Feng, L., 1998. Stock Movement and n-Dimensional Intertransaction Association Rules, *In Proceedings of 1998 SIGMOD Workshop, Research Issues on Data Mining and Knowledge Discovery,* Vol. 12, pp. 1-7.

Magdon-Ismail, M. and Atiya, A., 2004. *Maximum drawdown, Risk*, Vol. 17, No. 10, pp. 99-102.

Shen, L., Shen, H., and Cheng, L., 1999. New algorithms for efficient mining of association rules, *Information Sciences*, 118 (1–4), pp. 251–268.

Yao, H. and Hamilton H. J., 2006. Mining itemset utilities from transaction databases, *Data & Knowledge Engineering*, Vol. 59, pp. 603 – 626

Zhang D. and Zhaou L., 2004. Discovering golden nuggets: data mining in financial application, *IEEE transactions on Systems, Man, and Cybernetics – Part C: Application and Review,* Vol. 34(4), pp. 513-522.

Trade Station, 2011. *TradeStation Securities, Inc. for simulated trading.* http://www.tradestation.com/default_2.shtm