Stock Data Mining through Fuzzy Genetic Algorithm[1]

Longbing Cao, Chao Luo, Jiarui Ni, Dan Luo, Chengqi Zhang

Faculty of Information Technology, University of Technology, Sydney, Australia
{lbcao, jiarui, chengqi}@it.uts.edu.au

Abstract

Stock data mining such as financial pairs mining is useful for trading supports and market surveillance. Financial pairs mining targets mining pair relationships between financial entities such as stocks and markets. This paper introduces a fuzzy genetic algorithm framework and strategies for discovering pair relationship in stock data such as in high dimensional trading data by considering user preference. The developed techniques have a potential to mine pairs between stocks, between stock-trading rules, and between markets. Experiments in real stock data show that the proposed approach is useful for mining pairs helpful for real trading decision-support and market surveillance.

Keywords: stock data mining, fuzzy genetic algorithm

1. Introduction

Stock data mining has a potential to provide information for trading decision support and market surveillance. In stock data mining [3, 4], pairs mining targets discovering stock pairs from a series of stocks, a set of markets, or between instruments and trading rules in the market. Pairs mining is also promising for finding pair relationships between stocks and derivatives (namely stock-derivative pairs) lodged in an exchange. Pairs mined are helpful for traders to make smart trading decisions, or for market supervisors to monitor market integrity and abnormal trading behavior.

It is challenging to find out pairs of interest to business requirements in the real world. The following lists some of key challenges in mining pairs in domain-oriented business situations.

(i) The first is that pair relationship is usually hidden in high dimensional data. For instance, stock pairs are hidden in all stocks listed in an exchange, the number of listed stocks can be over 1,000.

(ii) The second challenge is that user preference and business needs are the drivers to generate pairs of interest to real user needs. Pairs must satisfy uncertain user requests often existing in real life. For instance, stock pairs are not steady in a market. To some users, a pairs of stock are highly correlated, while to others, the correlation is strictly checked by real requests such as beating market return.

(iii) The third challenge is to mine pairs crossing two different classes such as across two markets. Therefore, it is important to tackle the above factors in identifying and evaluating pairs in stock data.

This paper only discusses the issues in identifying pairs in high dimensional data. In this area, genetic algorithms (GA) [2] are widely used. However, GA itself is not good at dealing with domain-oriented business requests and user preference. To this end, this paper studies fuzzy genetic algorithms on top of traditional data mining techniques. Correlation analysis is used for mining pair relationship. It is further merged into fuzzy genetic algorithms. Fuzzy aggregation and fuzzy ranking are developed to handle the emerging issues in fuzzy GA. The proposed techniques have been used in mining stock pairs such as stock-stock pairs, stock-trading rule pairs, etc. For example, pairs trading strategy can be designed based on mined paired stocks. It supports to trade a basket of stocks to distribute potential trading and investment risk rather than putting all money on one instrument.

The organization of this paper is as follows. Section 2 discusses a fuzzy GA framework. In Section 3, fuzzy aggregation is introduced. We discuss fuzzy ranking in Section 4. An example of developing pairs trading strategy by utilizing mined stock pairs is demonstrated in Section 5. We conclude the paper in Section 6.

2. Fuzzy GA Integrating Fuzzy Set

By integrating genetic algorithm with fuzzy set [5], a fuzzy genetic algorithm [1] is a fuzzy set-coded genetic algorithm, where each individual (chromosome) is composed of a set of membership functions. In designing fuzzy genetic algorithms, issues in conventional genetic algorithms are put into

[1] This work is co-sponsored by Australian Research Council Discovery Grant (DP0667060), Centrelink research fund, UTS ECRG and Chancellor grants, and China Overseas Outstanding Talent Research Program of Chinese Academy of Sciences (06S3011S01).

fuzzy context and converted into fuzzy versions, for instance, fuzzy representation, fuzzy genetic operators, etc. In particular, fuzzy genetic algorithms need to consider the validation and ranking of the created fuzzy sets.

In mining pairs in financial markets, we develop the following fuzzy genetic algorithm framework. It deals with all basic issues such as initialization, selection, crossover, mutation and evaluation of stock pairs mining in fuzzy context. For initialization, all individuals are sampled randomly within the valid domain.

ALGORITHM 1: pseudo code for fuzzy genetic algorithm
Input: real number set X
Output: optimal fuzzy set Y for decision support
Procedure: FGA($\mu, X(t), \bar{X}(t), \bar{X}'(t), Y$)

```
//start with an initial time
t := 0;
//initialize a fuzzy random population of individuals X̄(t) by
fuzzifying the real number sets X(t) with proper membership
functions μ_X̄,
initialize
X̄(t) = {(x, μ_X̄(x)) | x ∈ X(t), μ_X̄ : X(t) → [0,1]};
//evaluate the fitness of all initial individuals of population
based on fuzzy evaluation
evaluate X̄(t);
//test for termination criterion
While (not done) do
    //increase the time counter
    t := t + 1;
    //select a fuzzy sub-population set X̄'(t) for offspring
production
    X̄'(t) := select X̄(t);
    //crossover the "genes" of the selected parents X̄'(t)
    crossover X̄'(t);
    //perturb the mated population stochastically
    mutate X̄'(t);
    //fuzzily evaluate its new fitness
    evaluate X̄'(t);
    //select the survivors Ȳ from actual fitness
    Ȳ := survive X̄(t), X̄'(t);
End
//fuzzily rank the survivors
rank Ȳ;
//defuzzify and export the final survivors
export Y;
```

Based on domain knowledge, we use *sharpe ratio* as the fitness function for all individuals in the pair population. It encloses information regarding risk and profit of a trading behavior. Furthermore, we fuzzify the real coded sharpe ratio into the interval [0,1] to get its fuzzy sets $\overline{SR}$. We use triangle piecewise linear membership function to fuzzify the universal sets. For instance, we specify ten levels of linguistic values, namely 1^{st}, 2^{nd}, …, 10^{th} from the lowest to the highest, for the fuzzy linguistic variable *sharpe ratio* $\overline{SR}$. Hereby we generate top $\bar{N}$ target objects, for instance the corresponding trading rules, $\bar{N}$ refers to those rules corresponding to the first N highest linguistic values.

In real coded genetic algorithms for pairs mining, its crossover can be in an arithmetic and/or multiple-point manner. We provide multi-point arbitrary crossover in a shuffling probability $p(0 \le p \le 1)$ of alleles on top of the top $\bar{N}$ selected sub-populations. On the other hand, the mutation is based on changing the original value stochastically by the mutation rate q $(0 \le q \le 1)$ either positively or negatively. Mutation operation is conducted on top of the shuffled sets $\bar{X}'(t)$ with the rate q around 0.03.

Optimized individuals emerge from the candidate population. They are possible optimal candidates for the final recommendation. In order to generate the final optimal list, special attention should be paid to the aggregation, evaluation and ranking of fuzzy functions and fuzzy sets. The following section discusses these issues.

3. Fuzzy Aggregation

One of the key objectives of fuzzy genetic algorithms is to recommend a set of optimal individuals *Y*. To this end, fuzzy aggregation and fuzzy ranking play an important role in generating the final survivors. The following procedures illustrate the process of finding out actionable trading rules highly correlated to given stocks in the market.

(i) The first is to mine and rank the in-depth trading rules [4] for a specific stock.
(ii) The second step is to detect and order the very appropriate stocks for a given trading rule.
(iii) We further aggregate these two lists through fuzzy aggregation rules to obtain a set of composite optimal stock-trading rule pairs.
(iv) Finally, we fuzzily rank the trading rule-stock pairs, and further defuzzify them to generate final outputs.

Let *sharpe ratio* be fitness function for the above first two steps. Suppose we build ten ascending linguistic values from 1^{st} to 10^{th}. To distinguish the two cases, as illustrated in Figure 1, we use fuzzy linguistic terms *a* to *j* and fuzzy values *A* to *J* to label the fuzzy sets for the optimal trading rules given a specific stock (we called rule sets), and for the appropriate stocks given a trading rule (called stock sets), respectively.

In practice, even though sharpe ratio is used as the fitness and similar linguistic measures are used for both rule set and stock set situations, the meaning of a specific corresponding linguistic term, say *b* and *B* in this case, may be highly varying. This means that we cannot aggregate the stock-rule pairs based on the

equal matching of two linguistic values from different sets. Instead, we develop the following solution to aggregate the two fuzzy groups.

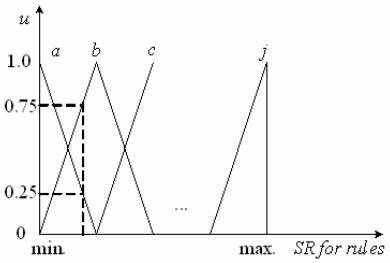

Fig. 1: Fuzzy trading rule set

Another fuzzy variable *rule-stock pair* (in short *pair*) is used to correlate close partners between the rule set and the stock set to aggregate the two groups. The *pair* has 19 linguistic values ascending from *1st*, 2nd to *19th*. Figure 2 defines its triangle fuzzy sets.

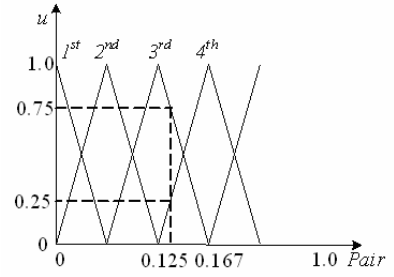

Fig. 2: Fuzzy set for trading rule-stock pairs

Further, the following fuzzy aggregation rules are defined to merge the fuzzy sets from different groups. For instance, if the rule set is *c* (i.e., 3rd), and the stock rule is *d* (4th), then the rule-stock pair is ranked as *6th* (3+4-1).

DEFINITION. *For the fuzzy rule set m-th, and the fuzzy stock set n-th, the rule-stock pair is (m+n-1)-th.*

Based on fuzzy aggregation rule, we can aggregate rule set and stock set into a universal ranking set by listing all possible candidate ranking. Table 1 illustrates the results of fuzzy aggregation and ranking of linguistic values *a*, *b*, *c* from the rule set and values *A*, *B*, *C* from the stock set, respectively. The fuzzy rule makes it possible to integrate the rule set and the stock set, and output the higher ranked rule-stock pairs as final survivors for optimal decision-making.

Table 1. Fuzzy aggregation and ranking of rule sets and stock sets

	A	*b*	*c*
A	*1st*	*2nd*	*3rd*
B	*2nd*	*3rd*	*4th*
C	*3rd*	*4th*	*5th*

4. Fuzzy Ranking

Fuzzy aggregation and ranking strategy is based on the fuzzification of fitness and membership functions. A rule in fuzzy set *c* or a stock in fuzzy set *D* is basically from fuzzy rather than crisp perspective. For instance, a trading rule could be classified into fuzzy set *b* with a membership grade μ=0.75 or set *a* with the grade μ=0.25. Similar thing exists for stock set, a stock could be segmented into fuzzy set *B* or *C* with the same grade μ=0.5. In this case, the outcome of the fuzzy aggregation and ranking could have four options, namely *2nd*, *3rd*, *4th* or *5th*.

It is necessary to manage the above uncertain situations in fuzzy aggregation and ranking. To this end, a ranking coefficient ρ based on moment defuzzification is introduced. m refers to the number of triggered linguistic values, l =1, 2, …, m corresponds to each triggered linguistic value. μ_l^R is the membership grade of No. l linguistic term relevant to the sharpe ratio of a rule. μ_l^S is the membership grade of No. l linguistic term corresponding to the sharpe ratio of a stock. η_l is the centroid of the No. l triggered linguistic value, it is calculated in terms of the moment and the area of each subdivision.

$$\rho = \sum_{l=1}^{m} \eta_l \mu_l^R \mu_l^S \Big/ \sum_{l=1}^{m} \mu_l^R \mu_l^S$$

ρ defuzzifies a fuzzy set returning a floating point that represents the fuzzy set. It actually measures how optimal a pair is. It deals with possible uncertainty when a rule-stock pair is aggregated. A real number can be obtained to measure a fuzzy rule-stock pair in a relatively crisp manner. For instance, we can calculate and get ρ =0.125 in the above example. As shown in Figure 2, this clearly indicates that this rule-stock pair is ranked as *3rd* fuzzy set since its membership grade is 0.75 which is larger than fuzzy set *4th* with grade 0.25.

5. Developing Unexpected Trading Evidences from Paired Stocks

The above fuzzy genetic algorithm techniques are used to mine financial pairs such as stock pairs and rule-stock pairs. Due to space limitation, here we only introduce an approach to discovering effective unexpected pairs trading evidence using fuzzy GA. We analyze correlations between stocks, and develop trading rules to trade the correlated stock pairs. Stock correlation analysis gets involved in analyzing market dynamics and microstructure. We design the following algorithms to discover highly correlated stocks in ASX by considering correlation coefficient and market factors. We find a dozen of unexpected stock pairs which are of surprise to traders. We then develop a pairs trading strategy to trade these pairs.

ALGORITHM 2: Method of discovering a pairs trading strategy based on stock correlation analysis

Input: a set of historical intraday orderbook transactions T, a set of stocks S, a coefficient threshold $coeff_0$, a sharpe ratio threshold sr_0, a return threshold r_0 ,

Output: a set of correlated stocks S_c, optimal distance d_0

Method:

1. Calculating the correlation coefficient ρ between two stocks A and B ($\geq coeff_0$);
2. Evaluating the actionability of the correlated stocks through cooperation with traders, considering market dynamics such as relationships to market sectors, volatility, liquidity and market index;
3. Generating actionable correlated stock list, e.g., if A and B are correlated both technically and from market dynamics perspective, then exists A-B pair;
4. Training the following pairs trading strategy (i.e., trade A and B alternatively) using fuzzy genetic algorithm:
   IF $P_A - \delta * P_B >= d_0$, THEN buy B and sell A
   IF $P_A - \varphi * P_B <= -d_0$, THEN sell B while buy A
   Where P_A and P_B are prices of A and B, d_0 is the distance factor, δ and φ are weights.
5. Optimizing parameter combination of d_0, δ and φ through fuzzy genetic algorithm in terms of fuzzified business metrics (say sharpe ratio and return) and user guide;
6. Testing the above-discovered pairs trading rules in out-of-sample data;
7. Fuzzily evaluating and ranking the stock pairs in terms of both technical and business metrics, balancing return, risk and the number of trading signals (success ratio);
8. Iteratively evaluating and refining the strategy by considering the relationships with volatility, liquidity and market index;
9. Exporting stock pair list and coupled combination among d_0, δ and φ as a pairs trading strategy.

In testing ASX Top 32 stocks from January 1997 to June 2002, we find 13 highly correlated stocks that are of surprise to traders. All 13 mined stocks come from different sectors. This finding means that pairs are not necessary from the same sector as presumed by traders and financial researchers. Originally traders didn't expect any relations between these stocks. Tests show that the return of trading on this pair is over 40% in ASX orderbook from 1 Jan 2002 to 18 Jun 2002 before considering transaction costs. This result is much more than the traders' expectation.

In real-world mining, it is very time-consuming to work out a trading evidence of interest to traders. We learn the lessons that the development of trading evidence actionable to real trading needs to involve domain knowledge and the existing constraints. For instance, Figure 4 further shows the impact of business factors – distance and weight on return and the number of triggered signals. Further, correlation relationship between stocks and the combination of the above four factors interesting to trading cannot just be determined by technical measures such as correlation coefficient. They are also highly affected by stock movement such as volatility and liquidity. High volatility improves return while high liquidity balances the market impact on return.

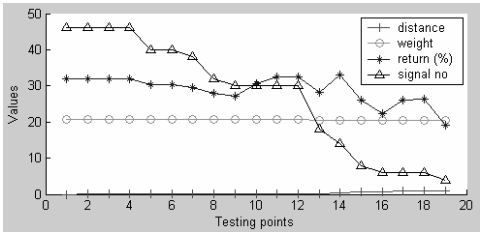

Fig. 4: Relation between d_0, ϕ, *return* and *signal number* (Stock: CBA and GMF, Data: ASX orderbook from 1 Jan to 20 Jun 2000)

6. Conclusions

In this paper, we introduce some of our research in developing techniques for pairs mining. In particular, we develop a fuzzy genetic algorithm framework, which combine genetic algorithm with fuzzy set. Special strategies are developed for fuzzy aggregation and ranking. The proposed approach is useful for handling high dimensional data with consideration of domain-oriented requirements and user preference. We illustrate the research of developing unexpected trading evidences based on finding correlated stocks. Our further work is on combining parallel genetic algorithms with the developed fuzzy genetic algorithms.

7. References

[1] J. Buckley, Y. Hayashi, "Fuzzy genetic algorithm and applications," *Fuzzy Sets and Systems*, 61: 129-136, 1994

[2] L. Davis, (Ed.). *Handbook of genetic algorithms*. New York: Van Nostrand Reinhold, 1991

[3] B. Kovalerchuk, E. Vityaev, *Data Mining in Finance: Advances in Relational and Hybrid Methods*, Kluwer Academic, 2000

[4] L. Lin, L. Cao, "Mining In-Depth Patterns in Stock Market," *Int. J. Intelligent System Technologies and Applications*, 2006

[5] L. Zadeh, "Fuzzy sets," *Information and Control*, 83: 338–353, 1965