Using Genetic Algorithms to Discover Selection Criteria for Resolving Contradictory Solutions Returned by CBR

Brent Stephens

Master's Oral Defense

May 20, 2005

Case Based Reasoning

- Problem Solving Method
  - Results easily understood by users
- Direct application of experience to new problems
  - Case Base
  - Similarity Metric
  - Adaptation

CBR for Classification

- Solution is classification
- Simpler version
- No adaptation
- Learning by retention

Domain

- Real world domains
  - Corporate database
  - Large and Redundant
  - Unstructured and Error prone

- BNSF Railroad
  - Shipping data
  - Correcting unclassified cases by assigning a billing code
    - User Errors
    - Domain Shifts
    - Cyclical billing
  - Existing Rule Based System was inadequate

CBR Properties

- Weighted matching
- Minimum normalized similarity threshold
- Resulting case set
  - All solutions match
  - Contradictory solutions returned

Limitation of CBR in this Domain

- Contradictory solutions retrieved
  - No method available from experts to select correct solution
- Options
  - Maintenance of Case Base
    - Eliminate redundant or contradictory solutions
    - Expensive because of the volume of new cases
    - May require lots of work by operator
  - Improve Similarity Metric
    - Inaccuracy or incompleteness of expert matching methods
- Experts recommended looking at other qualities of set of cases retrieved

Problem Significance

- CBR ability to deal with contradictory solution
- Better apply CBR to real world domains
- Better emulate expert knowledge that is difficult to apply
- Replace workers in doing tedious, boring work
- Unique in that it applies properties of the returned cases rather than features

Solution

- Selection criteria for contradictory cases
- Basic formulas used to derive solution
- Use Genetic Algorithms to learn formulas

Implementation

- Use CBR to retrieve cases
  - Features and weights given by experts
- Frequency and recency
  - Features of returned cases recommended by experts but no method of applying them is given
- Discover formulas to determine significance of both
- Use Genetic Algorithms to determine formulas

Frequency and Recency

- Frequency
  - Percentage of cases with a common solution
- Recency
  - How long before new case did retrieved case occur
  - Maximum age is learned by GA

Scoring

- Frequency or recency score fed into formula
- Result multiplied by CBR score
- Scores for a solution are summed within formula
- Total scores for formula are normalized
- Highest scoring solution is selected

Example Formulas

- Step $\left\lfloor \frac{frequency}{1/\alpha} \right\rfloor \times \beta$
- Exponential $\alpha \times e^{-(\beta(1-recency))} + \gamma$
- Linear $\alpha \times frequency - \beta$

Additional Formulas

- Most Recent
- Most Frequent
- K-Nearest Neighbor

Combining Scores

- Weighting for each formula learned by GA
- Score generated for each solution by each formula
- Scores normalized
- Final score for a solution generated by summing weighted formula scores

GA Properties

- Generation Size - 1000
- Number of generations -1000
- Mutation – 1%
- Crossover Mating – 99%
- Succeeding generation creation
- Variable Representations

Formula Learning Procedure

- Training set – 10 sets of 50 cases
  - Chromosome converted to variables
  - Set of training cases evaluated
  - Fitness formula applied to results
  - Next generation created
  - Switch to next training set
- Repeat for all 6 formulas
- Repeat at each minimum similarity

Fitness Formulas

- Fitness Formula 1
  - Percentage of cases correctly classified
- Fitness Formula 2
  - Percentage of cases correctly classified
  - Difference in score when correctly classified
  - Difference in score when incorrectly classified

Resultant Formula Example

- Fitness Formula 2
- Minimum Similarity .98

- Step function for frequency
  - *cutoff date =16*

$$= \left\lceil \frac{recency}{1/7} \right\rceil \times 0.02$$

Combination weight learning

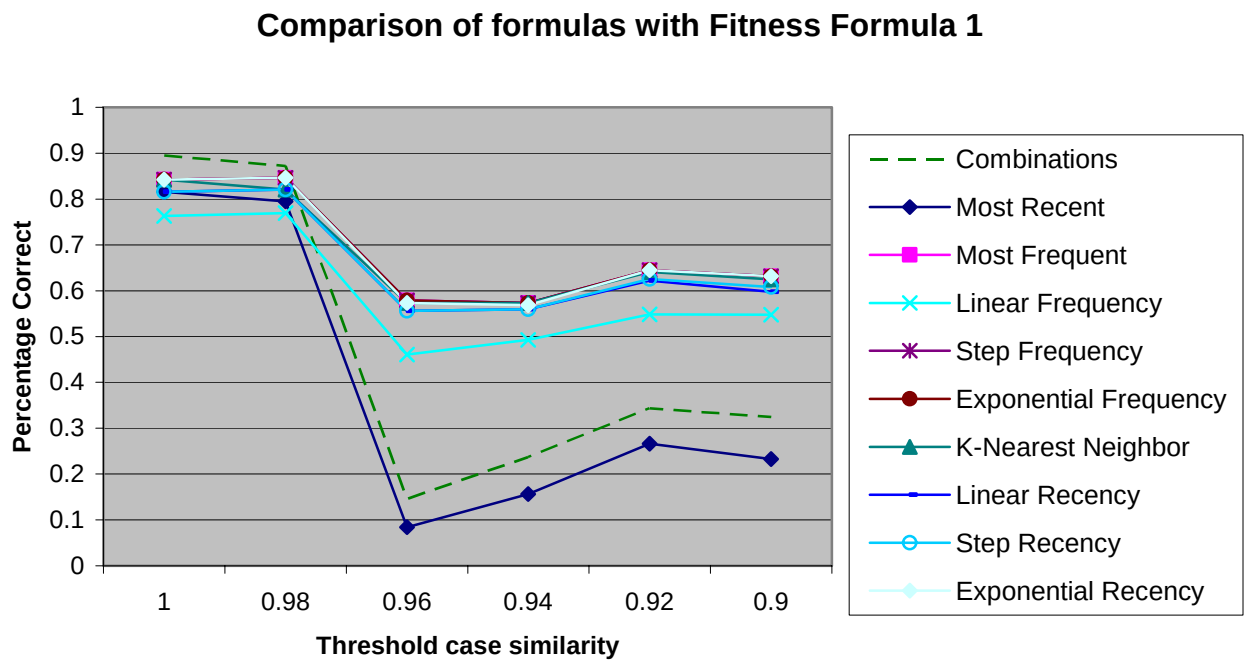
- After formula learning is completed
- Same fitness formulas used

$$= \omega_1 f_1 + \omega_2 f_2 + \omega_3 f_3 + \ldots$$

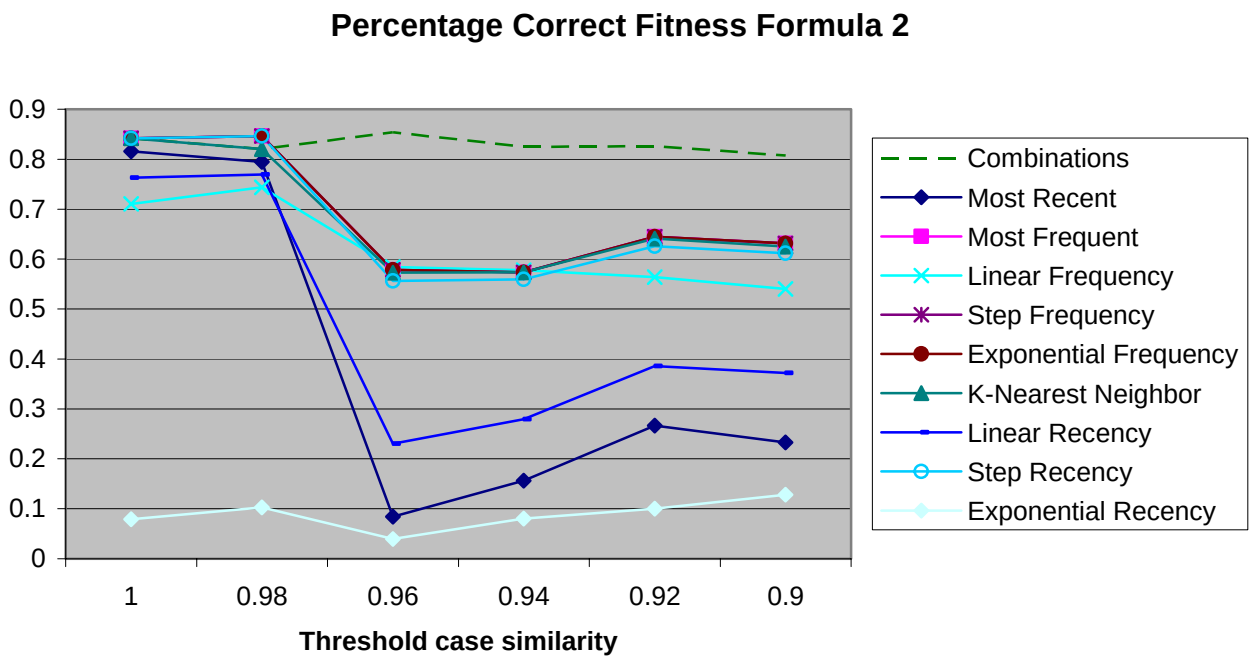
Testing procedure

- Test set – 500 cases
- CBR Matching
- Formulas Evaluated
- Formula scores combined
- Correctness checked for individuals formulas and combined formulas

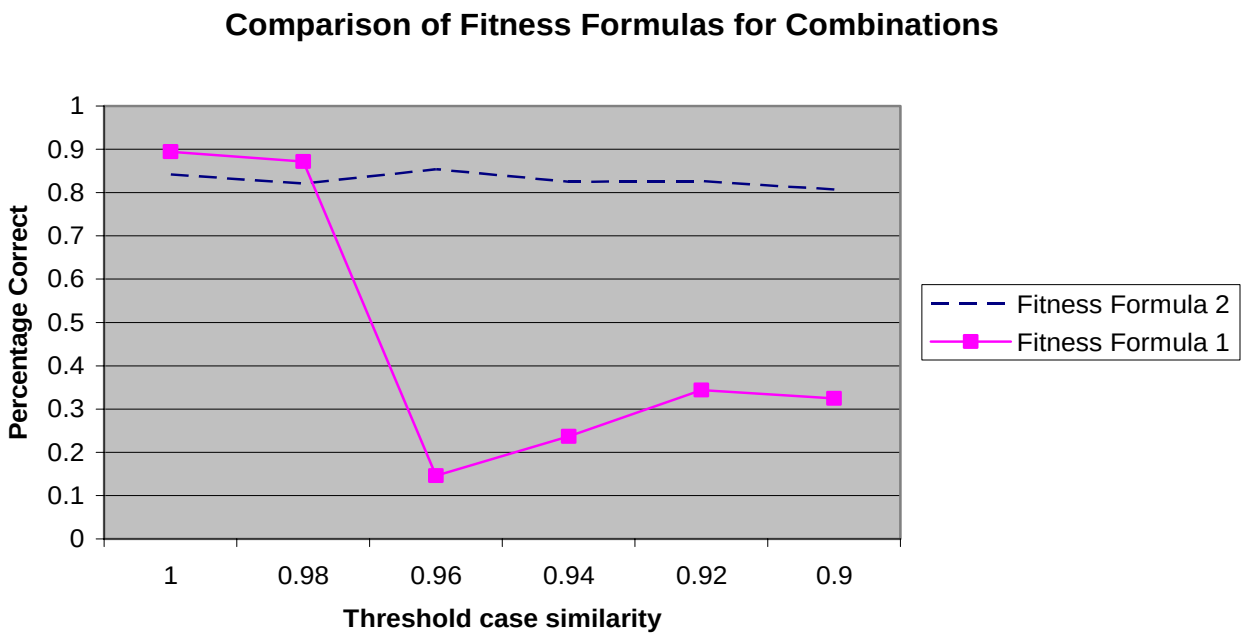
Formula 1 Classification Rate



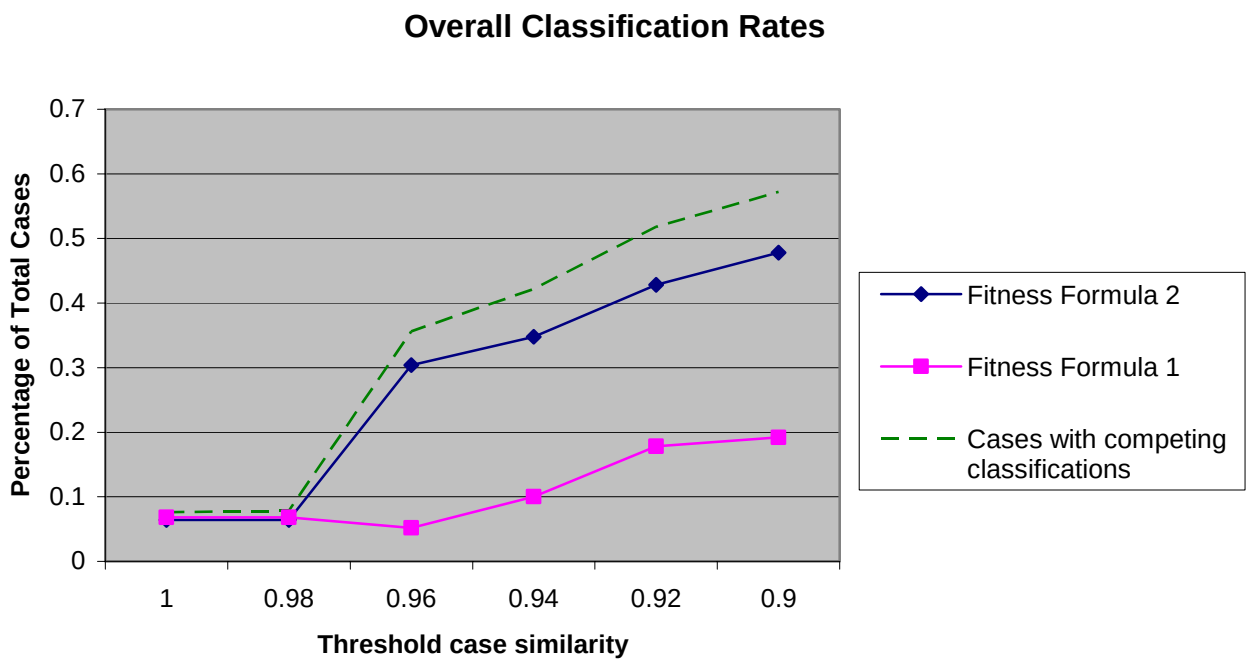
Formula 2 Classification Rate



Fitness Formula Accuracy



Overall Classification Rate



Meaning

- GA trained formulas show significant improvement over traditional selection methods
- Combined solution outperformed individual formulas

Conclusions

- Improve performance of CBR using GAs
- Selection of features and formulas appropriate to domain
- Fitness method significantly affects performance

Conclusions

- Combining results improved performance
- Applicable in domains where expert knowledge is incomplete or inaccurate