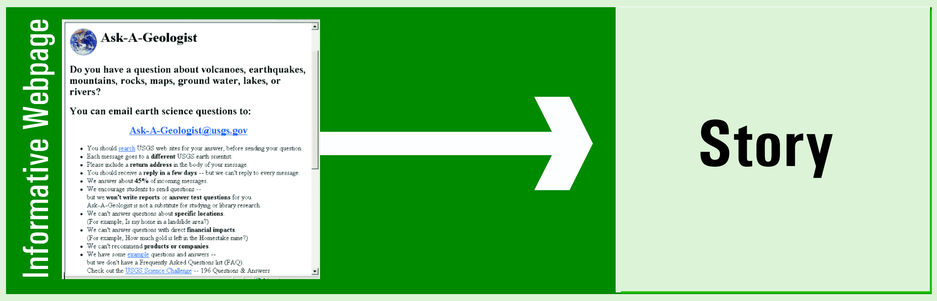
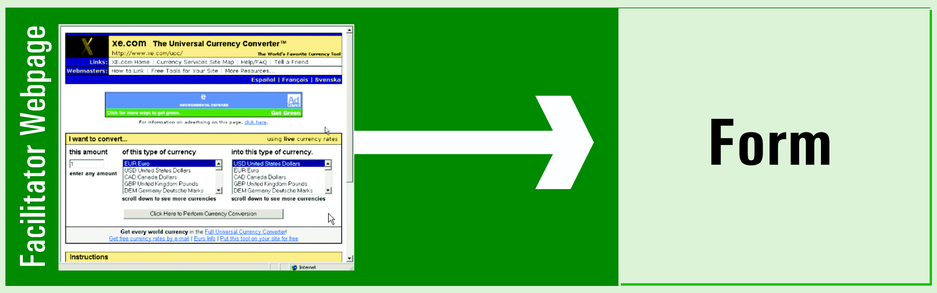
Solving Problems Two at a Time

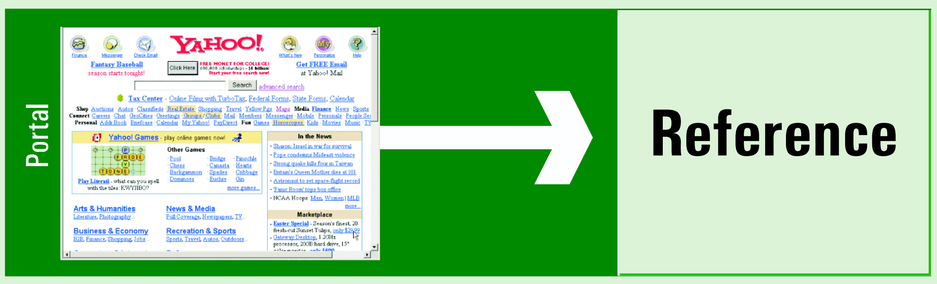
Hassan Alam, Fuad Rahman and Yuliya Tarnikova

Classification of Web Pages using a

Generic Pair-Wise Multiple Classifier System







Web database

400 samples in that database, hree-fourth is used for training and one one-fourth is used for testing.

The classes are: Reference pages (pages primarily of links and referrals, sites such as Yahoo!), story pages (pages primarily of textual content, sites such as CNN), Form pages (pages with large interactive forms, sites that deal with banking and transactions).

A Pair-wise Classifier: Support Vector Machine (SVM)

Structural Risk Minimization

Vapnik-Chervonenkis (VC) Dimension
- -Property of set of functions $\{f(\alpha)\}$
- -Maximum number of training points that can be shattered by $\{f(\alpha)\}$
- -Ex R^N's VC dimension of the set of oriented lines

$$h = n + 1$$

VC Theory provides bounds on the test error, which depend on both empirical risk and capacity of function class

$$R(\alpha) \le R_{emp}(\alpha) + \phi\left(\frac{h}{l}, \frac{\log(\eta)}{l}\right)$$

$$\phi\left(\frac{h}{l}, \frac{\log(\eta)}{l}\right) = \sqrt{\frac{h(\log\frac{2l}{h}+1) - \log(\frac{\eta}{4})}{l}}$$

Features

-Ratio of text to links
-Largest chunk of text
-Maximum size of text per column
-Ratio of largest continuous chunk of text to links
-Ratio of text to embedded links
-Ratio between contiguous text and contiguous links
-Ratio of contiguous chunk of text to total text
-Number of images
-Images with and without links
-Average contiguous text and link between images
-Ratio of non-repeating links to repeating links
-Boldness
-Underlines
-Highlighting
-Headline or other tags
-Links in navigation columns
-Number of active form-related tags
-Ratio of forms to links
-Ratio of forms to text and ratio of text chunks that are preceded by a link.

Accuracy

Story 93.94%
Forms 81.81%
Reference 87.88%

Configuration: ensemble of pair-wise classifiers resolving one-to-many conflicts before resolving one-to-one conflicts.

Overall 87.88%

Configuration: ensemble of pair-wise classifiers resolving one-to-one conflicts before resolving one-to-many conflicts.

Overall 86.87%

Initial Configuration

Conflict in 1st layer: 8%

The 2nd layer ensemble corrected 25% of the wrong classifications delivered by the first layer ensemble.

In the other 75% cases, the second layer ensemble supported the (either correct or wrong) decision of the first layer ensemble.

In no cases a correct decision was discarded.

Alternate Configuration

-Ensemble of pair-wise classifiers resolving one-to-one conflicts before resolving one-to-many conflicts.

-The second layer was activated in 4% of the cases
In 25% of the cases the incorrect decisions were corrected

-In 25% of the cases correct decisions were discarded in favor of incorrect decisions

-In 50% of the cases, first layer decisions were supported. The accuracy of this configuration was 86.87%, lower than that achieved with the first configuration

Decision

The original configuration is much more stable, especially in terms of the false positive and false negative conflict resolution.

Conclusions

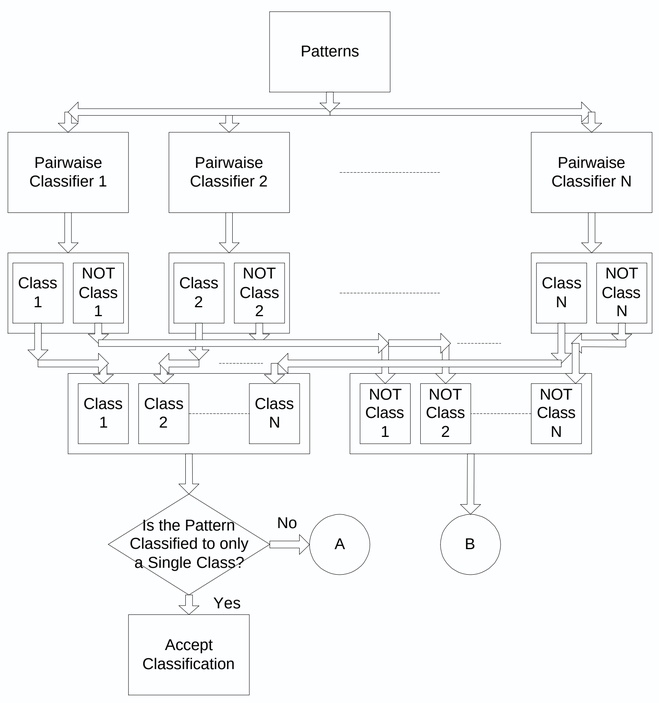
-A new generic solution to the web classification problem.

-A generic pair-wise multiple classifier system was presented

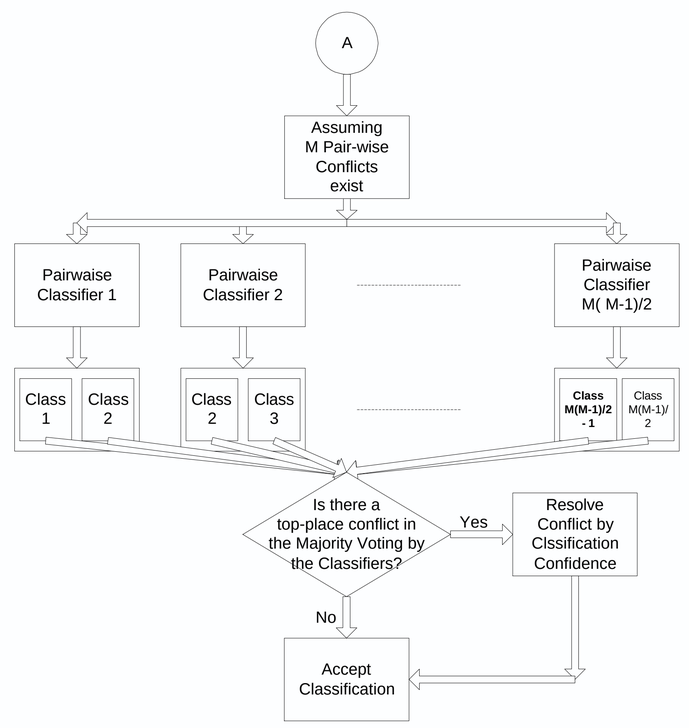
-This MCS configuration made up of individual pair-wise classifiers is able to provide very high accuracy in a very difficult problem domain.

-It is also shown how decision flow through the system can be tracked and how individual layers of this two-layer ensemble can be separately controlled.

The generic pair-wise MCS



Re-evaluation of patterns with specific ensemble of classifiers



Re-evaluation of patterns with specific ensemble of classifiers