An Ensemble of Case-Based Classifiers for High-Dimensional Biological Domains

Niloofar Arshadi[a] and Igor Jurisica[a,b]

[a] Department of Computer Science, University of Toronto, 10 King's College Road, Toronto, Ontario M5S 3G4, Canada;

[b] Ontario Cancer Institute, Division of Cancer Informatics, Toronto Medical Discovery Tower, Life Sciences Discovery Centre, 101 College Street, Toronto, Ontario M5G 1L7

ABSTRACT

In this paper, we propose the *mixture of experts for case-based reasoning* (MOE4CBR), a method that combines an ensemble of case-based classifiers with spectral clustering and logistic regression. We demonstrate the improvement achieved by applying the method to a computational framework of a CBR system called *TA3*. We evaluate the system with different number of CBR classifiers (experts) on two publicly available mass spectrometry data sets on ovarian cancer. Our proposed method improves the classification accuracy of *TA3* from 90.1% to 99.2% on the ovarian data set 8-7-02, and from 79.1% to 96.3% on the ovarian data set 4-3-02.

Keywords: Case-based reasoning, clustering, feature selection, mixture of experts

1. INTRODUCTION

Case-based reasoning (CBR) has been successfully applied to a wide range of applications including classification, diagnosis, planning, configuration, and decision-support.[1] CBR can produce good quality solutions in weak theory domains such as molecular biology, where the number and the complexity of the rules affecting the problem are very large, there is not enough knowledge for formal knowledge representation, and our domain understanding evolves over time.[2] Protein expression profiling using mass spectrometry is a recent method for profiling cancer cases to measure thousands of elements in a few microliters of serum,[3] and also an example of high-dimensional molecular biology domain. The data obtained are mass-to-charge ratios (m/z values) for individual proteins. Mass spectrometry data sets are represented by two-dimensional matrices, where each row contains the mass-to-charge values for proteins (known as biomarkers) for cancer and control (normal) samples. In addition, clinical information is used to label and further describe individual samples.

Using principles of case medicine for diagnosis and prognosis, CBR naturally fits this application domain. However, (ultra) high-dimensionality of mass spectrometry data sets (tens of thousands of biomarkers with only few hundreds of samples) poses a challenge that needs to be addressed. One solution is to combine CBR classifiers with other machine learning techniques to improve the prediction accuracy and overcome the "curse of dimensionality".

We propose an ensemble of CBR systems, called the *mixture of experts* (MOE) to predict the classification label of an unseen data (query). A gating network calculates the weighted average of votes provided by each expert. We apply spectral clustering[4] to cluster the data set (case-base) into k groups. Each cluster is considered as a case-base for the k CBR experts, and the gating network learns how to combine the responses provided by each expert. The performance of each CBR expert is further improved by using feature selection techniques. We use logistic regression[5] to select a subset of features in each cluster. We demonstrate the improvement achieved by applying our method to a specific implementation of a CBR system, called *TA3*.[6] *TA3* is a computational framework for CBR based on a modified nearest neighbor technique.

The rest of the paper is organized as follows. Section 2 introduces our proposed method. Section 3 introduces the *TA3* CBR system, which is used as a framework for evaluating MOE4CBR. In Section 4, we demonstrate the experimental results of the proposed method on two publicly available ovarian data sets.

Further author information: N.A.: E-mail: niloofar@cs.toronto.edu
I.J.: E-mail: juris@ai.utoronto.ca

2. THE MIXTURE OF EXPERTS FOR CASE-BASED REASONING (MOE4CBR) METHOD

The goal of our method is to improve the prediction accuracy of CBR classifiers using the mixture of experts. The performance of each expert in MOE4CBR is improved using clustering and feature selection techniques. The MOE4CBR method comprises of the following components:

Clustering: Of the many clustering approaches that have been proposed, only some algorithms are suitable for domains with large number of features and a small number of samples. Our earlier comparison of self-organizing maps (SOMs)[7] with k-means clustering[8] and spectral clustering[4] suggests that spectral clustering outperforms k-means clustering and SOMs.[9]

Feature Selection: Feature selection (FS) improves the quality of data by removing redundant and irrelevant features, i.e., those features whose values do not have meaningful relationships to their labels, and whose removal improves the prediction accuracy of the classifier. We compared Fisher criterion with standard t-test[10] and logistic regression model[5] when applied as a feature selection, logistic regression outperforms the other two feature selection techniques in terms of accuracy and classification labels.[9] In our method, feature selection is applied in two steps: (1) on the whole case-base, FS selects a subset of features; (2) after grouping the case-base into k clusters, FS selects a subset of features in each group.

Mixture of Experts: The mixture of experts approach is based on the idea that each expert classifies samples separately, and individual responses are combined by the gating network to provide a final classification label.[5] It consists of two parts: experts and the gating network. For an unseen query case, each expert of CBR retrieves l similar cases from its case-base (l can be chosen by the user). It should be noted that experts do not share their case-bases, rather the case-base of each expert is obtained by clustering the whole case-base into k non-overlapping clusters (k can be chosen by the user or estimated by other analysis). In other words, the initial case-base is clustered into k groups, and then feature selection techniques are applied to each case-base to extract more "informative" biomarkers.

After retrieving l similar cases from the case-base, the expert applies the weighting vote algorithm[9] to predict the class label of the query case, i.e., performs weighted case adaptation. More precisely, let $\{C_1, \ldots, C_k\}$ denote the clusters (or the k case-bases of our k experts), x the unseen data, y a class label, S_j the number of similar cases that belong to C_j, and T_j the number of similar cases with class label y that belong to C_j, $Pr(Y = y|C_j, x)$ is then computed as $\frac{T_j}{S_j}$.

The gating network combines the classification labels predicted by each expert. First, it assigns weights to each expert – represented by g_j using CBR, where $1 \leq j \leq k$. Briefly, g_j represents the probability that the unseen data x belongs to the case-base of the j^{th} expert. The initial case-base is used in retrieving m similar cases (m can be chosen by the user). In order to compute g_j that can be shown as $Pr(C_j|x)$, we perform the following steps. Let R_j represent the number of similar cases to x belonging to C_j (the case-base of the j^{th} expert), g_j then is calculated by dividing R_j by m. Finally, in order to combine the responses of k experts, following formulas are used[5]:

$$Pr(Y = y|x) = \sum_{j=1}^{k} g_j \times Pr(Y = y|C_j, x), \tag{1}$$

with the constraint that:

$$\sum_{j=1}^{k} g_j = \sum_{j=1}^{k} Pr(C_j|x) = 1, \tag{2}$$

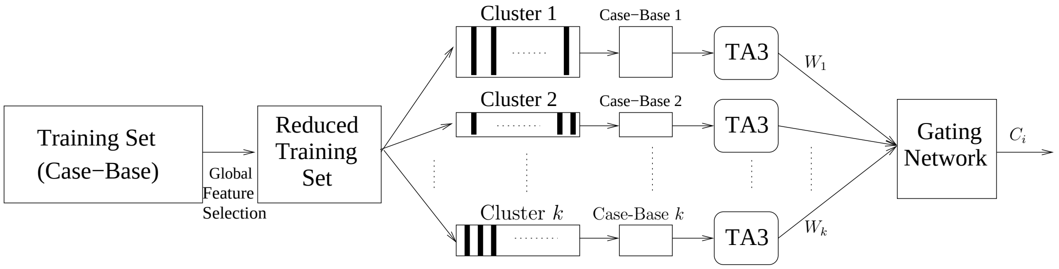

Figure 1. Mixture of Experts for Case-Based Reasoning: Training set is grouped into k clusters, and after selecting a subset of features for each group (shown with vertical bars), each group will be used as a case-base for the k CBR experts. The gating network combines the responses provided by each *TA3* expert considering the weights of each expert (weights are shown on the arrows connecting *TA3* experts to the gating network.

3. AN INTRODUCTION TO TA3 CASE-BASED REASONING SYSTEM

We used the *TA3* CBR system as a framework to evaluate our method. The *TA3* system has been applied successfully to biology domains such as *in vitro fertilization* (IVF)[11] and protein crystal growth.[12] This section briefly describes the system.

Case Representation in TA3: In classification tasks, each case has at least two components: problem description and class. The problem description characterizes the problem and the class gives a solution to a given problem. In this domain, a case corresponds to the sample profile represented by a vector of m/z values.

Case Retrieval in TA3: The retrieval component is based on a modified nearest neighbor matching[13] (See[6] for more details). Similarity in *TA3* is determined as a closeness of values for attributes defined in the *context*. Context can be seen as a view or an interpretation of a case, where only a subset of attributes are considered relevant.

Case Adaptation in TA3: The adaptation process in CBR manipulates the solution of the retrieved case to better fit the query. We adopt distance-weighted nearest neighbor[14] to determine the classification label of the query based on the labels of similar retrieved cases (See [9] for more details).

4. EXPERIMENTAL RESULTS

The experiments have been performed on the following mass spectrometry data sets provided online at the National Institutes of Health and Food and Drug administration Clinical Proteomics Program Databank. *

Ovarian data set 8-7-02: Ovarian data set 8-7-02 comprises 162 mass spectra from ovarian cancer patients and 91 individuals without cancer (control group) with 15,154 mass-to-charge ratios (m/z values) measured in each serum.

Ovarian data set 4-3-02: Ovarian data set 4-3-02 contains spectra from 100 patients with ovarian cancer and 116 individuals without cancer (control group). The serum mass spectrum for each subject consists of 15,154 mass-to-charge ratios.

Table 1 depicts the results of applying MOE4CBR to our two ovarian data sets with different number of experts. When there is a tie, the *TA3* classifier cannot decide on the label; resulting cases are categorized as "undecided" in the Table. We used 10-fold cross-validation for validation, and the Table shows the average over the 10 folds. In each iteration, MOE4CBR was trained using 9 folds, and was tested on the remaining fold, i.e., the test set was completely unseen until the test time, and clustering and feature selection techniques were applied only to the training set.

When there is only one expert – *TA3* classifier – the case-base does not split into groups, and the size of the case-base is reduced by selecting 15 biomarkers out of 15,154 biomarkers. For the ovarian data set 8-7-02,

*`http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`

the minimum classification error is achieved when the number of experts equals 2, while for the ovarian 4-3-02, the minimum classification error is achieved with 3 experts (Table 1). Table 2 compares the case where only a single instance of *TA3* (without applying feature selection or clustering techniques) is applied to classify our two data sets with the case that MOE4CBR classifies the data sets.

Table 1. Accuracy of MOE4CBR with different number of experts(shown with n) on ovarian data sets. In all experiments, 15 biomarkers were selected by logistic regression, and the whole case-base was clustered into smaller groups using spectral clustering.

Ovarian Data Set 8-7-02			
	$n = 1$	$n = 2$	$n = 3$
Accuracy	98%	99.2%	98.8%
Error	2%	0.8%	1.2%
Undecided	0%	0%	0%
Ovarian Data Set 4-3-02			
	$n = 1$	$n = 2$	$n = 3$
Accuracy	94.9%	95.4%	96.3%
Error	4.6%	4.6%	3.7%
Undecided	0.5%	0%	0%

Table 2. Accuracy of MOE4CBR compared to the case where a single instance of *TA3* without applying feature selection or clustering techniques is used for classification.

Ovarian Data Set 8-7-02			
Method	Accuracy	Error	Undecided
Single TA3	90.1%	9.1%	0.8%
MOE4CBR	99.2%	0.8%	0%
Ovarian Data Set 4-3-02			
Method	Accuracy	Error	Undecided
Single TA3	79.1%	18.6%	2.3%
MOE4CBR	96.3%	3.7%	0%

These two ovarian data sets have been previously analyzed.[15,16] Sorace et al.[16] evaluate their extracted rules for selecting biomarkers on data set 8-7-02 when it is randomly split into training and test data. They achieve 100% sensitivity and 100% *specificity*. The rules are extracted in an "ad hoc" way, and might not be applicable to other similar data sets. Ovarian data set 4-3-02 has also been analyzed by Zhu et al.,[15] and they achieve 100% specificity and 100% sensitivity. They also show that their 18 selected biomarkers classifies ovarian data set 8-7-02 into cancer and control samples with 100% sensitivity. It has been recently reported that their results cannot be replicated and the overall best performance achieved using the proposed 18 markers is 98.42%.[17] Our results are not comparable with those two studies, since we used 10-fold cross-validation, while they split the data set randomly into training and test set.

5. CONCLUSIONS

Molecular biology is a natural application domain for CBR systems, since CBR systems can perform remarkably well on complex and poorly formalized domains. Removing "non-informative" features from the case-base of each member classifier helps overcome the "curse of dimensionality".

In this paper, we proposed the mixture of experts for case-based reasoning (MOE4CBR) method, where an ensemble of CBR systems is integrated with clustering and feature selection to improve the prediction accuracy of the *TA3* classifier. Spectral clustering groups samples, and each group is used as a case-base for each of the k experts of CBR. To improve the accuracy of each expert, logistic regression is applied to select a subset of features that can better predict class labels. We also showed that our proposed method improves the prediction accuracy of the *TA3* case-based reasoning system on two public ovarian data sets. In future, we will evaluate the method on other high-dimensional life sciences domains.

ACKNOWLEDGMENTS

This work is supported by the National Science and Engineering Research Council of Canada (NSERC Grant 203833-02) and IBM Faculty Partnership Award to IJ. The authors are grateful to Patrick Rogers, who implemented the current version of *TA3*.

REFERENCES

1. M. Lenz, B. Bartsch-Sporl, H. Burkhard, and S. Wess, eds., *Case-Based Reasoning: experiences, lessons, and future directions*, Springer, 1998.
2. I. Jurisica and J. Glasgow, “Application of case-based reasoning in molecular biology,” *Artificial Intelligence Magazine, Special issue on Bioinformatics* **25**(1), pp. 85–95, 2004.
3. E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, “Use of proteomic patterns in serum to identify ovarian cancer,” *Lancet* **359**(9306), pp. 572–577, 2002.
4. A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in *Advances in Neural Information Processing Systems 14*, Z. G. G. Dieterich, S. Becker, ed., MIT Press, 2002.
5. T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer, 2001.
6. I. Jurisica, J. Glasgow, and J. Mylopoulos, “Incremental iterative retrieval and browsing for efficient conversational CBR systems,” *International Journal of Applied Intelligence* **12**(3), pp. 251–268, 2000.
7. T. Kohonen, *Self-Organizing Maps*, Springer, 1995.
8. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kauffmann Publishers, 2000.
9. N. Arshadi and I. Jurisica, “Data mining for case-based reasoning in high-dimensional biological domains,” *IEEE Transactions on Knowledge and Data Engineering* **17**(8), pp. 1127–1137, 2005.
10. J. Devore, *Probability and statistics for engineering and the sciences*, Duxbury Press, 1995.
11. I. Jurisica, J. Mylopoulos, J. Glasgow, H. Shapiro, and R. F. Casper, “Case-based reasoning in IVF: prediction and knowledge mining,” *Artificial Intelligence in Medicine* **12**, pp. 1–24, 1998.
12. I. Jurisica, P. Rogers, J. Glasgow, S. Fortier, J. Luft, J. Wolfley, M. Bianca, D. Weeks, and G. DeTitta, “Intelligent decision support for protein crystal growth,” *IBM Systems Journal* **40**(2), pp. 394–409, 2001.
13. D. Wettschereck and T. Dietterich, “An experimental comparison of the nearest neighbor and nearest hyperrectangle algorithms,” *Machine Learning* **19**(1), pp. 5–27, 1995.
14. T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
15. W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm, and J. S. Kovach, “Detection of cancer-specific markers amid massive mass spectral data,” *Proceedings of the National Academy of Sciences of the United States of America* **100**(25), pp. 14666–14671, 2003.
16. J. M. Sorace and M. Zhan, “A data review and re-assessment of ovarian cancer serum proteomic profiling,” *BMC Bioinformatics* **4**(24), pp. 14666–14671, 2003. available at http://www.biomedcentral.com/1471-2105/4/24.
17. K. A. Baggerly, J. S. Morris, S. R. Edmonson, and K. R. Coombes, “Signal in noise: Evaluating reported reproducibility of serum proteomic tests for ovarian cancer,” *Journal of National Cancer Institute* **97**(4), pp. 307–309, 2005.