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

Niloofar Arshadi[1] and Igor Jurisica[1,2]

[1] Department of Computer Science, University of Toronto,
10 King's College Road, Toronto, Ontario M5S 3G4, Canada
niloofar@cs.toronto.edu
[2] Ontario Cancer Institute, Princess Margaret Hospital,
University Health Network, Division of Cancer Informatics,
610 University Avenue, Toronto, Ontario M5G 2M9, Canada
juris@ai.utoronto.ca

Abstract. It has been shown that an ensemble of classifiers increases the accuracy compared to the member classifiers provided they are diverse. One way to produce this diversity is to base the classifiers on different case-bases. In this paper, we propose the mixture of experts for case-based reasoning (MOE4CBR), where clustering techniques are applied to cluster the case-base into k groups, and each cluster is used as a case-base for our k CBR classifiers. To further improve the prediction accuracy, each CBR classifier applies feature selection techniques to select a subset of features. Therefore, depending on the cases of each case-base, we would have different subsets of features for member classifiers.
Our proposed method is applicable to any CBR system; however, in this paper, we demonstrate the improvement achieved by applying the method to a computational framework of a CBR system called *TA3*. We evaluated the system on two publicly available data sets on mass-to-charge intensities for two ovarian data sets with different number of clusters. The highest classification accuracy is achieved with three and two clusters for the ovarian data set 8-7-02 and data set 4-3-02, respectively. The proposed ensemble method improves the classification accuracy of *TA3* from 90% to 99.2% on the ovarian data set 8-7-02, and from 79.2% to 95.4% on the ovarian data set 4-3-02. We also evaluate how individual components in MOE4CBR contribute to accuracy improvement, and we show that feature selection is the most important component followed by the ensemble of classifiers and clustering.

1 Introduction

Case-based reasoning (CBR) has been successfully applied to a wide range of applications such as 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].

H. Muñoz-Avila and F. Ricci (Eds.): ICCBR 2005, LNAI 3620, pp. 21–34, 2005.

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) of varying intensities. Mass spectrometry data sets are represented by two-dimensional matrices, where each row contains the mass-to-charge intensities (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". Ensembles improve the accuracy of CBR classifiers [4, 5]; however, since k-nearest neighbor (kNN) and CBR classifiers are categorized under *stable* classifiers, having diverse classifiers is essential to improve the accuracy [6]. Stable classifiers are stable with respect to small changes in the training data.

One way to have diversity for stable classifiers is to select different subsets of features for each classifier [4, 5]. In this paper, in addition to selecting a different subset of features for each member classifier, we cluster the case-base into smaller groups. Data clustering means to group items (data points or attributes) into classes such that items within a cluster are similar to one another and dissimilar to items in other clusters. Thus, by grouping the whole case-base into smaller clusters, different classifiers would have different case-bases.

The goal of feature selection is to identify "informative" features among thousands of available features, i.e., relevant features that improve CBR performance for a given reasoning task. For mass spectrometry data sets, mining a subset of features that distinguishes between cancer and normal samples can play an important role in disease pathology and drug discovery. Early detection of cancer can reduce mortality, and identified biomarkers may also be useful drug discovery targets that may lead to new therapeutical approaches. Moreover, removing "non-informative" features helps overcome the "curse of dimensionality", and improves the prediction accuracy of classifiers.

Our hypothesis can be summarized as follows. Combining an ensemble of CBR classifiers with feature selection and clustering techniques not only helps overcome the "curse of dimensionality", but also leads to diverse classifiers, which is essential for improving the accuracy of ensembles. Our approach has three main components: (1) an ensemble of CBR systems, (2) clustering, and (3) feature selection. In principle, any CBR system, clustering, and feature selection algorithm can be used. However, the choice has to satisfy our performance criteria, which is to maximize prediction accuracy, and be applicable to high-dimensional domains.

We use 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 [7] 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 [8] to select a subset of features in each cluster.

Although the proposed method is applicable to any CBR system, we demonstrate the improvement achieved by applying it to a specific implementation of a CBR system, called *TA3* [9]. *TA3* is a computational framework for CBR based on a modified NN technique and employs a variable context, a similarity-based retrieval algorithm, and a flexible representation language.

The rest of the paper is organized as follows. Section 2 reviews ensembles, clustering, and feature selection techniques. In Section 3, we present MOE4CBR, a method that uses the mixture of CBR experts to classify high-dimensional data sets. Section 4 introduces the *TA3* CBR system, which is used as a framework for evaluating MOE4CBR. In Section 5, we demonstrate the experimental results of the proposed method on two publicly-available ovarian data sets.

2 Related Work

Ensembles improve the stability and accuracy of classifiers if there is diversity in the classifiers [6, 5]. If small changes in training data produces quite different models and thus different predictions, the learner is called an unstable learner [5]. Neural networks and decision trees are examples of unstable learners. For such classifiers, diversity can be achieved if classifiers are trained on different subsets of training data. However, since lazy learners such as kNN and CBR classifiers are relatively stable in the face of changes in training data [6], other sources of diversity must be employed. One way of achieving diversity is to consider a different subset of features for each classifier. Ricci and Aha [4] create various NN classifiers, each one considers a different subset of features and then their predictions are combined using error-correcting output codes (ECOCs). Cunningham and Zenobi [5] show that an ensemble of kNN classifiers based on different feature subsets can classify more accurately than a single kNN classifier based on the best feature subset available.

Clustering and *feature selection* techniques have been applied to many domains including high-dimensional biological domains [10, 11, 12]. Clustering groups samples (cases) into partitions such that samples within a cluster are similar to one another and dissimilar to samples in other clusters. Clustering techniques can be categorized into *partitional* and *hierarchical* methods [13]. Partitional-based clustering techniques attempt to break a data set into k clusters such that each cluster optimizes a given criterion, e.g., minimizes the sum of squared distance from the mean within each cluster. Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones (agglomerative approach), or by splitting larger clusters (divisive approach).

Clustering and feature selection techniques have been applied to CBR systems as well. Yang and Wu [14] propose a method that groups the case-base into smaller case-bases, and then each case-base is maintained individually. They use *density-based* clustering technique [15] in which a cluster is a region with a higher density of points than its surrounding region.

Shiu and Yeung [16] cluster the case-base into smaller partitions and select representative cases for each cluster in order to reduce the size of case-base. In their clustering approach, the similarity matrix of cases is formed, and two cases will be placed in the same cluster if their weighted Euclidean distance is smaller than a predetermined threshold.

Smyth and McKenna [17] cluster the case-base by finding the *related set* of each case. The related set of each case is the union of the set of cases that can be solved by this case and the set of cases that this case can solve. Two cases will be in the same cluster if the intersection of their related sets is not empty. Common problem types are typically represented by large and densely packed clusters, while smaller clusters, or even lone cases, generally represent more unusual problem types. Those cases that do not make critical competence contribution could be deleted. In their case-base editing approach, the size of case-base is minimized, while the range of problems that can be solved remains unchanged.

Feature selection techniques are classified into *filter* and *wrapper* methods [18]. The filter approach selects feature subsets that are independent of the induction algorithm, while the wrapper approach evaluates the subset of features using the inducer, itself. Aha and Bankert [19] discuss how using filter and wrapper techniques improve the classification accuracy of case-based classifiers on the cloud data set with 204 features and a few thousand data points. Their results show that a wrapper feature selection method (called BEAM) applied to an NN classifier improves its prediction accuracy by $\sim$20%.

3 The 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. Using the results of our earlier performance evaluation [20], we selected spectral clustering [7] for clustering the case-base, and the logistic regression model [8] as a filter feature selection for the *TA3* classifier. Given a labeled training data set, predicting labels of the unseen data (query), is performed in two steps: (1) each CBR experts retrieves l similar cases from its respective (non-overlapping) case-base; (2) the class label for the query is predicted by assigning weights to each expert. We discuss the process in details in the next section.

3.1 Mixture of Experts

The mixture of experts approach is based on the idea that each expert classifies data points (cases) separately, and individual responses are combined by the

gating network to provide a final classification label [21]. A general idea of the mixture of experts approach is depicted in Figure 1. In the first step, 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).

After retrieving l similar cases from the case-base, the expert applies the weighting vote algorithm (see Section 4.3) 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 $\frac{T_j}{S_j}$.

We use CBR to assign weights to each expert – represented by g_j, $1 \leq j \leq k$. Briefly, g_j represents the probability that the unseen data x belongs to the case-base of the j^{th} expert. More precisely, in order to compute g_j that can be shown as $Pr(C_j|x)$ as well, we perform the following steps. Let m represent the number of similar cases retrieved from the whole initial case-base by the gating network (m can be chosen by the user), R_j 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 [8]:

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

with the constraint that:

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

As Figure 2 depicts, the MOE4CBR method has two main steps: First, the case-base of each expert is formed by clustering the case-base into k groups. Second, each case-base selects a subset of features "locally". Each of the k obtained sets is considered as a case-base for our k experts of CBR. We use Equations 1 and 2 to combine the responses of the k experts. Each expert applies the *TA3* classifier to decide on the class label, and the gating network uses *TA3* to assign weights (represented by g_j) to each classifier as explained above.

3.2 Clustering

Of the many clustering approaches that have been proposed, only some of the algorithms are suitable for domains with (ultra) high number of features and a low number of samples. The two widely used clustering approaches in (ultra) high-dimensional DNA microarrays [22, 23] are k-means clustering [13] and

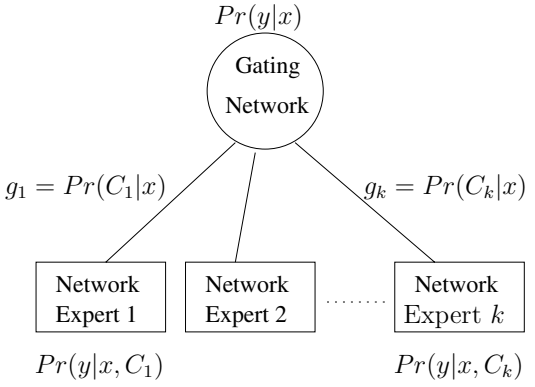

Fig. 1. Mixture of Experts: Terminal nodes are experts, and the non-terminal node is the gating network. The gating network returns the probability that the query x belongs to class Y

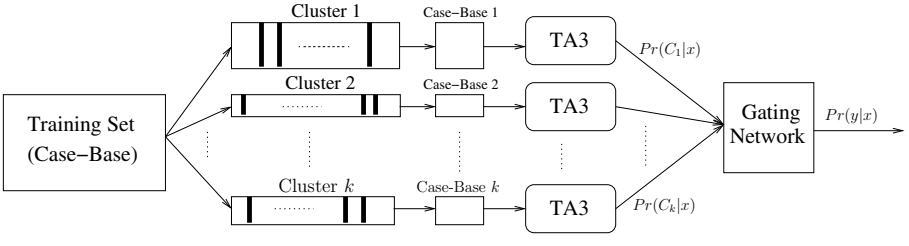

Fig. 2. Mixture of Experts for Case-Based Reasoning: Training set is grouped into k clusters, and after selecting a subset of features for each group (shown with vertical bars), each group will be used as a case-base for the k CBR experts. The gating network combines the responses provided by each *TA3* expert considering the weights of each expert (weights are shown on the arrows connecting *TA3* experts to the gating network)

self-organizing maps (SOMs) [24]. Our earlier evaluation suggests that spectral clustering [7] outperforms k-means clustering and SOMs [20].

Spectral clustering is based on the approach where data points are mapped to a new space, prior to being clustered. More precisely, first, a matrix X holding the Euclidean distance between any two data points (i.e., a transformation of the *affinity* matrix) is formed. Second, matrix Y is formed from X by stacking the k eigenvectors associated with the k largest eigenvalues of matrix X in columns. Each row of Y is treated as a point in $\mathcal{R}^k$ and is clustered into k clusters using k-means algorithm , where k represents the number of clusters and is set by the user. In the next step, data point s_i is assigned to cluster j if and only if row i of the matrix X was assigned to cluster j, where $1 \leq i \leq N$, $1 \leq j \leq k$, and N is the number of data points. This clustering technique has been successfully used in many applications, including computer vision and VLSI [7].

3.3 Feature Selection

The goal of feature selection is to improve 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.

Fisher's criterion and standard t-test are two statistical methods that have been successfully applied to feature selection problem in (ultra) high-dimensional data sets [25]. In order to select a suitable feature selection approach, we evaluated the performance of Fisher's criterion, t-test, and the *logistic regression* model [8] when used in a CBR classifier [20]. We applied the three feature selection techniques to the *TA3* classifier, and measured the improvement in *accuracy* and *classification error*. Accuracy measures the number of correctly classified cases, and classification error counts the number of misclassified cases. Based on our evaluation, logistic regression as a feature selection method outperforms Fisher and standard t-test techniques [26].

Assuming that classifier x is the logistic of a linear function of the feature vector, for two classes, the logistic regression model has the following form:

$$Pr(y = 0|x, w) = \frac{1}{1 + e^{-w^T x}}, \tag{3}$$

where w is a $p + 1$ column vector of weights, and p is the number of features [8]. Logistic regression has been successfully applied to classifying (ultra) high-dimensional microarrays [27]. However, we use the logistic regression classifier as a filter feature selection method. In order to select a subset of features (genes), the logistic regression classifier is trained using the above Equation on the training set, and features corresponding to the highest ranking magnitude of weights are selected. The data sets are normalized such that all features (regressor variables) have the same mean and variance.

4 An Introduction to the TA3 Case-Based Reasoning System

Although our method can be applied to any CBR 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) [28] and protein crystal growth [29]. This section briefly describes the system.

4.1 Case Representation in *TA3*

A case C corresponds to a real world situation, represented as a finite set of attribute/value pairs [28]. Using the information about the usefulness of individual attributes and information about their properties, attributes are grouped into two or more Telos-style categories [30]. 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. Additional categories can be used to group attributes into separate equivalence partitions, and treating each partition separately during case retrieval.

4.2 Case Retrieval in *TA3*

The retrieval component is based on a modified NN matching [31]. Its modification includes: (1) grouping attributes into categories of different priorities so that different preferences and constraints can be used for individual categories during query relaxation; (2) using an explicit context (i.e., set of attribute and attribute value constraints) during similarity assessment; (3) using an efficient query relaxation algorithm based on incremental context transformations [9].

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. By selecting only certain features for matching and imposing constraints on feature values, a context allows for controlling what can and what cannot be considered as a partial match: all (and only) cases that satisfy the specified constraints for the context are considered similar and are relevant with respect to the context.

4.3 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 [32] to determine the classification label of the query based on the labels of similar retrieved cases. Let $x_1, ..., x_k$ denote the k cases retrieved from the case-base that are similar to the query x_q. In order to predict the label of x_q shown with $\hat{f}(x_q)$, following equations are used [32]:

$$\hat{f}(x_q) \leftarrow \underset{v \in V}{argmax} \sum_{i=1}^{k} \omega_i \delta(v, f(x_i)),$$

where

$$\omega_i \equiv \frac{1}{d(x_q, x_i)^2},$$

and V is the finite set of class labels $\{v_1, ..., v_s\}$, $f(x_i)$ the class label of case x_i, and $\delta(a, b) = 1$ if $a = b$ and $\delta(a, b) = 0$ otherwise.

5 Experimental Results

Here we demonstrate the results of applying the MOE4CBR method to the *TA3* classifier. In [33], we showed MOE4CBR improves the prediction accuracy of high-dimensional microarrays. In this study, we show the improvement in the classification accuracy of two publicly mass spectrometry data sets by applying MOE4CBR. Also, we experiment MOE4CBR with different number of experts and evaluate its components separately.

5.1 Data Sets

The experiments have been performed on the following mass spectrometry data sets. The two mass spectrometry data sets [34, 35] discussed in this paper, are both provided online at the National Institutes of Health and Food and Drug administration Clinical Proteomics Program Databank. [1]

1. *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.
2. *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.

These two ovarian data sets have been previously analyzed [34, 35, 26, 20]. Sorace et al. [34] evaluate their extracted rules for selecting biomarkers on data set 8-7-02 when it is randomly split into training and test data. Although they achieve 100% sensitivity and 100% specificity, our results are not comparable, as they evaluated their method on randomly selected training and test sets, while we used 10-fold cross-validation. Also, their 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. [35]. They achieve 100% specificity and 100% sensitivity. Our results are not comparable, since we used 10-fold cross-validation, while they split the data set randomly into training and test set. Furthermore, it had been recently reported that their results cannot be replicated and the overall best performance achieved using the proposed 18 markers is 98.42% [36].

Similarly, these two ovarian data sets have been analyzed using a *TA3* classifier combined with logistic regression [26]. This approach resulted in 98% accuracy and 2% error for the ovarian data set 8-7-02, and 95.4% accuracy and 4.6% error for the ovarian data set 4-3-02, evaluated using 10-fold cross-validation.

Each of the studies have selected a different set of "informative" biomarkers, and further biological validation, which is beyond the scope of this paper, will be able to determine which list of biomarkers is clinically more "informative" for diagnosis or drug discovery of ovarian cancer samples.

5.2 Evaluating MOE4CBR with Different Number of Experts

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

[1] `http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`

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				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$
Accuracy	98%±2.8%	98.4%±2%	99.2%±1.6%	96.4%±2.9%
Error	2%	1.2%	0.8%	2.8%
Undecided	0%	0.4%	0%	0.8%
Ovarian Data Set 4-3-02				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$
Accuracy	95.4%±4.3%	95.4%±4.8%	94.9%±5%	90.3%±4.9%
Error	4.6%	4.1%	5.1%	7.8%
Undecided	0%	0.5%	0%	1.9%

was tested on the remaining fold, i.e., the test set was quite 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, the minimum classification error is achieved when the number of experts equals 3, while for the ovarian 4-3-02, the minimum classification error is realized with 2 experts (Table 1).

5.3 Evaluating Components of MOE4CBR

We used 10-fold cross-validation to evaluate our proposed method in terms of accuracy, classification error, and "undecided" rate, and the results are averaged over 10 folds. We evaluated the components of MOE4CBR as follows:

- This is our base line, where a single instance of *TA3* classifies the query case without being integrated with any FS or clustering technique, and only a single classifier predicts the label.
- In order to evaluate the FS component, we use logistic regression to select 15 biomarkers out of 15,154 biomarkers, and then we apply *TA3* as a CBR classifier.
- In order to evaluate the clustering component, we split the case-base randomly into two groups, and use MOE4CBR to classify the query case. In this case, the number of experts equals 2, logistic regression selects 15 biomarkers, and the results are averaged over 5 iterations.
- Finally, we apply MOE4CBR when logistic regression as a filter FS method selects 15 biomarkers, and spectral clustering groups the case-base into two clusters (i.e., there are only two experts).

As the Table 2 shows, the FS component contributes the most in improving the accuracy of the classifier, while spectral clustering has the least contribution. As is typically found in most studies, kNN and CBR classifiers are

Table 2. Accuracy of MOE4CBR with different components

Ovarian Data Set 8-7-02			
Method	Accuracy	Error	Undecided
Single TA3	90%	9.2%	0.8%
TA3 with LR	98%	2%	0%
MOE4CBR with LR and RC	97.4%	2.6%	0%
MOE4CBR with LR and SC	98.4%	1.2%	0.4%
Ovarian Data Set 4-3-02			
Method	Accuracy	Error	Undecided
Single TA3	79.2%	18.5%	2.3%
TA3 with LR	95.4%	4.6%	0%
MOE4CBR with LR and RC	94.6%	5.2%	0.2%
MOE4CBR with LR and SC	95.4%	4.1%	0.5%

Note. LR: Logistic Regression; RC: Random Clustering; SC: Spectral Clustering

very "sensitive" to the selected features and the "curse of dimensionality" problem. Therefore, removing "non-informative" features helps improve the accuracy. On the other hand, although spectral clustering outperforms k-means and self-organizing maps in terms of precision, recall, and Dunn's index [20], it still does not perform much better than random clustering. This can be due to the ultra high-dimensionality of data sets. Applying FS techniques before clustering may help improve the performance of clustering techniques.

6 Conclusions

Molecular biology is a natural application domain for CBR systems, since CBR systems can perform remarkably well on complex and poorly formalized domains. Although high dimensionality poses a challege and reduces system performance, the classification accuracy improves by using an ensemble of classifiers. Also, 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.

Although we have used a specific implementation of a CBR system, our results are applicable in general. Generality of our solution is also not degraded by the application domains, since many other life sciences problem domains are characterized by (ultra) high-dimensionality and a low number of samples.

Further investigation may take additional advantage of Telos-style categories in *TA3* for classification tasks. The system may also benefit from new clustering approaches, and other feature selection approaches such as wrapper and hybrid approaches.

Acknowledgments

This work is supported by IBM CAS fellowship to NA, and 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] Lenz, M., Bartsch-Sporl, B., Burkhard, H., Wess, S., eds.: Case-Based Reasoning: experiences, lessons, and future directions. Springer (1998)
[2] Jurisica, I., Glasgow, J.: Application of case-based reasoning in molecular biology. Artificial Intelligence Magazine, Special issue on Bioinformatics **25(1)** (2004) 85–95
[3] Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., Liotta, L.A.: Use of proteomic patterns in serum to identify ovarian cancer. Lancet **359(9306)** (2002) 572–577
[4] Ricci, F., Aha, D.W.: Error-correcting output codes for local learners. In Nedellec, C., Rouveirol, C., eds.: Proceedings of the 10th European Conference on Machine Learning, Springer (1998) 280–291
[5] Cunningham, P., Zenobi, G.: Case representation issues for case-based reasoning from ensemble research. In Aha, D.W., Watson, I., eds.: Case-Based Reasoning Research and Development:4th International Conference on Case-Based Reasoning, Springer (2001) 146–157
[6] Breiman, L.: Bagging predictors. Machine Learning **24** (1996) 123–140
[7] Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In G. Dieterich, S. Becker, Z.G., ed.: Advances in Neural Information Processing Systems 14, MIT Press (2002)
[8] Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer (2001)
[9] Jurisica, I., Glasgow, J., Mylopoulos, J.: Incremental iterative retrieval and browsing for efficient conversational CBR systems. International Journal of Applied Intelligence **12(3)** (2000) 251–268
[10] Xing, E.P.: Feature selection in microarray analysis. In Berrar, D., Dubitzky, W., Granzow, M., eds.: A practical approach to Microarray data analysis. Kluwer Academic publishers (2003) 110–131
[11] Quackenbush, J.: Computational analysis of microarray data. Nat Rev Genet **2** (2001) 418–427
[12] Molla, M., Waddell, M., Page, D., Shavlik, J.: Using machine learning to design and interpret gene-expression microarrays. AI Magazine **25** (2004) 23–44
[13] Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kauffmann Publishers (2000)

[14] Yang, Q., Wu, J.: Keep it simple: a case-base maintenance policy based on clustering and information theory. In Hamilton, H., ed.: Advances in Artificial Intelligence, In Proceedings of the 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, Montreal, Canada, Springer (2000) 102–114

[15] Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd international Conference on knowledge discovery and data mining, Portland, OR, USA, AAAI Press (1996) 226–231

[16] Shiu, S.C., Yeung, D.S.: Transferring case knowledge to adaptation knowledge: An approach for case-base maintenance. Computational Intelligence **17** (2001) 295–314

[17] Smyth, B., McKenna, E.: Building compact competent case-bases. In Althoff, K.D., Bergmann, R., Branting, K., eds.: Proceedings of the 3rd International Conference on Case-Based Reasoning Research and Development (ICCBR-99), Seeon Monastery, Germany, Springer (1999) 329–342

[18] John, G., Kohavi, R., Pfleger, K.: Irrelevant features and the subset selection problem. In: Machine Learning: Proceedings of the eleventh international conference, Morgan Kauffmann (1994) 121–129

[19] Aha, D.W., Bankert, R.: Feature selection for case-based classification of cloud types: an empirical comparison. In Aha, D.W., ed.: Proceedings of the AAAI-94 workshop on Case-Based Reasoning, Menlo Park, CA: AAAI Press (1994) 106–112

[20] Arshadi, N., Jurisica, I.: Data mining for case-based reasoning in high-dimensional biological domains. IEEE Transactions on Knowledge and Data Engineering (2005) To appear

[21] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixture of local experts. Neural Computation **3** (1991) 79–87

[22] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science **286** (1999) 531–537

[23] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Dmitrovsky, E., Lander, E., Golub, T.: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. In: Proceedings of the National Academy of Science of the United States of America. Volume 96(6). (1999) 2907–2912

[24] Kohonen, T.: Self-Organizing Maps. Springer (1995)

[25] Jaeger, J., Sengupta, B., Ruzzo, W.: Improved gene selection for classification of microarrays. In: Pacific Symposium on Biocomputing. (2003) 8:53–64

[26] Arshadi, N., Jurisica, I.: Feature selection for improving case-based classifiers on high-dimensional data sets. In: FLAIRS 2005 - The 18th International FLAIRS Conference, AAAI Press (2005) To appear

[27] Xing, E.P., Jordan, M.L., Karp, R.M.: Feature selection for high-dimensional genomic microarray data. In Brodley, C.E., Danyluk, A.P., eds.: Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, Morgan Kauffmann (2001) 601–608

[28] Jurisica, I., Mylopoulos, J., Glasgow, J., Shapiro, H., Casper, R.F.: Case-based reasoning in IVF: prediction and knowledge mining. Artificial Intelligence in Medicine **12** (1998) 1–24

[29] Jurisica, I., Rogers, P., Glasgow, J., Fortier, S., Luft, J., Wolfley, J., Bianca, M., Weeks, D., DeTitta, G.: Intelligent decision support for protein crystal growth. IBM Systems Journal **40(2)** (2001) 394–409

[30] Mylopoulos, J., Borgida, A., Jarke, M., Koubarakis, M.: Telos: Representing knowledge about information systems. ACM Transactions on Information Systems **8(4)** (1990) 325–362

[31] Wettschereck, D., Dietterich, T.: An experimental comparison of the nearest neighbor and nearest hyperrectangle algorithms. Machine Learning **19(1)** (1995) 5–27

[32] Mitchell, T.M.: Machine Learning. McGraw-Hill (1997)

[33] Arshadi, N., Jurisica, I.: Maintaining case-based reasoning systems: a machine learning approach. In Funk, P., González-Calero, P.A., eds.: Advances in Case-Based Reasoning: 7th European Conference, Springer (2004) 17–31

[34] Sorace, J.M., Zhan, M.: A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics **4:24** (2003) 14666–14671 available at `http://www.biomedcentral.com/1471-2105/4/24`.

[35] Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J., Kovach, J.S.: 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)** (2003) 14666–14671

[36] Baggerly, K.A., Morris, J.S., Edmonson, S.R., Coombes, K.R.: Signal in noise: Evaluating reported reproducibility of serum proteomic tests for ovarian cancer. Journal of National Cancer Institute **97(4)** (2005) 307–309