Feature Selection for Improving Case-Based Classifiers on High-Dimensional Data Sets

Niloofar Arshadi
Department of Computer Science,
University of Toronto,
10 King's College Road,
Toronto, Ontario M5S 3G4, Canada
niloofar@cs.toronto.edu

Igor Jurisica
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

Case-based reasoning (CBR) is a suitable paradigm for class discovery in molecular biology, where the rules that define the domain knowledge are difficult to obtain, and there is not sufficient knowledge for formal knowledge representation. To extend the capabilities of this paradigm, we propose *logistic regression for CBR* (LR4CBR), a method that uses logistic regression as a feature selection (FS) method for CBR systems. Our method not only improves the prediction accuracy of CBR classifiers in biomedical domains, but also selects a subset of features that have meaningful relationships with their class labels.

In this paper, we introduce two methods to rank features for logistic regression. We show that using logistic regression as a filter FS method outperforms other FS techniques, such as Fisher and t-test, which have been widely used in analyzing biological data sets. The FS methods are combined with a computational framework for a CBR system called $TA3$. We also evaluate the method on two mass spectrometry data sets, and show that the prediction accuracy of $TA3$ improves from 90% to 98% and from 79.2% to 95.4%. Finally, we compare our list of discovered biomarkers with the lists of selected biomarkers from other studies for the mass spectrometry data sets, and show the overlapping biomarkers.

Introduction

The case-based reasoning (CBR) approach is particularly suitable for applications in the life sciences , where we lack sufficient knowledge for formal representation or parameter estimation (Jurisica & Glasgow 2004). Two examples are gene and protein expression profiling using microarrays and mass spectrometry. Microarrays are used in biological and medical domains to provide profiles of diseased and normal tissues, in order to increase our understanding of biological processes, disease origin, state, type, and progression at a molecular level. These profiles will play a crucial role in information-based medicine in the future by enabling earlier and more accurate diagnosis, prognosis, and treatment planning.

Microarray data sets are represented by an $N \times M$ matrix of real values that represent gene expression values, where M is the number of genes used to profile N samples, and they are labeled using clinical profiles (or phenotypes). Another recent method for profiling biological samples such as cancer samples is mass spectrometry, used to measure thousands of elements in a few microliters of serum (Petricoin *et al.* 2002). The data obtained are mass-to-charge ratios (m/z values) of varying intensities. Mass spectrometry data sets, similar to microarray 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.

CBR has been applied to a wide range of tasks, such as classification, diagnosis, planning, configuration, and decision support (Lenz *et al.* 1998; Leake 1996). It does not rely on statistical assumptions and it has intuitive appeal because of its similarity to the human analogical reasoning in problem-solving. However, CBR classifiers, similarly to other classifiers, can suffer from the "curse of dimensionality" that occurs in (ultra-) high-dimensional domains with tens of thousands of attributes [1] and only a few hundreds of samples. Feature selection (FS) techniques help overcome the problem by selecting "informative" features among thousands of available features, i.e., those features that improve CBR performance for a given reasoning task, and thus are biologically meaningful. For example, in microarray data sets, "informative" features comprise genes with expression patterns that have meaningful biological relationships to the classification labels of samples (analogously, they could represent sample vectors that have meaningful biological relationship to the classification labels of genes).

Feature selection techniques have been successfully combined with CBR systems (Aha & Bankert 1994; Arshadi & Jurisica 2004). For both microarray and mass spectrometry data sets, mining a subset of features that distinguishes between cancer and normal samples (or other phenotypes) can play an important role in disease pathology and drug discovery. Early detection of cancer can reduce mortality, and identified markers may also be useful drug discovery targets that may lead to new therapeutical approaches.

In this paper, we propose *LR4CBR* (logistic regression for

[1] In this paper, we use attributes and features interchangeably, unless specified otherwise.

CBR) — a method that selects a subset of features using the logistic regression model for our $TA3$ CBR classifier. $TA3$ is a computational framework for CBR based on a modified nearest-neighbor technique that employs a variable context, a similarity-based retrieval algorithm, and a flexible representation language (Jurisica, Glasgow, & Mylopoulos 2000). We demonstrate the improvement in accuracy achieved by applying our method to two publicly available mass spectrometry data sets.

The paper is organized as follows. First, we introduce Fisher and t-test methods — the two FS methods widely used for analyzing (ultra-) high-dimensional biological data sets (Jaeger, Sengupta, & Ruzzo 2003; Zhu *et al.* 2003; Wu *et al.* 2003; Baggerly, Morris, & Coombes 2004). We then discuss logistic regression for FS. The logistic regression model has been used for classifying microarray data sets (Xing, Jordan, & Karp 2001); however, in this paper we apply it as a filter FS method. Our experiments reveal that logistic regression combined with our CBR classifier achieves higher accuracy than the two other FS methods for (ultra-) high-dimensional data sets. Then, we evaluate the proposed method on two publicly available mass spectrometry data sets. Finally, we present a subset of biomarkers that are differentially expressed in the sera of ovarian cancer patients, which after additional biological validation may prove to be useful for ovarian cancer diagnosis.

The Logistic Regression for Case-Based Reasoning (LR4CBR) Method

FS algorithms can be classified as either *filter* or *wrapper* methods (Kohavi & John 1997). The main difference between the two types is the use of the final classifier to evaluate the subset of features in the wrapper approaches, while filter methods do not use it.

In this paper, we focus on filter FS techniques. We present the results of a comparison of Fisher's criterion, t-test, and the logistic regression model (Hastie, Tibshirani, & Friedman 2001) with a CBR classifier. We applied the three FS techniques to the $TA3$ classifier, and measured its improvement in *accuracy* and *classification error*. Accuracy measures the number of correctly classified data points, and classification error counts the number of misclassified data points. We use both measures, as our classification system also supports the "undecided" label.

- The **Fisher's criterion score** is defined as $\frac{(m_1-m_2)^2}{(v_1+v_2)}$, where m_i and v_i are the mean and variance of the given feature in class i.
- The **Standard t-test** is defined as $\frac{|m_1-m_2|}{\sqrt{(\frac{v}{n_1}+\frac{v}{n_2})}}$, where m_i is the mean of the given feature in class i, n_i is the number of examples in class i, and v is the pooled variance across both classes (Devore 1995). The score is reported as a negative $\log_{10}$ p-value. The greater the score is, the more significant the difference between the means.
- **Logistic Regression** has been successfully applied to classifying (ultra-) high-dimensional microarrays (Xing, Jordan, & Karp 2001). However, we use the logistic regression model as a filter FS method. For two classes, the logistic regression model has the following form (Hastie, Tibshirani, & Friedman 2001):

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

where w is a $p + 1$ column vector of weights, and p is the number of features. We use maximum likelihood to estimate w:

$$l(w) = \sum_{i=1}^{N} logPr(y = y_i|x_i, w), \quad (2)$$

where x_i is the i^{th} training example augmented with a constant 1, y_i is the binary response for the i^{th} training example, and N is the number of training examples. To solve the above equation, we use stochastic gradient ascent:

$$w^{(t+1)} = w^{(t)} + \rho(y_n - \mu_n^{(t)})x_n, \quad (3)$$

where $w^{(t)}$ is the parameter vector at the t^{th} iteration, $\mu_n^{(t)} \equiv 1/(1 + exp(w_t^T x_n))$ and ρ is a step size, which is determined experimentally.

We compared two feature ranking criteria for logistic regression.

– The first criterion is often applied to linear classifiers (Mukherjee 2003). It trains the regression classifier using Equation 3, and selects features corresponding to the highest ranking magnitude of weights.

– LeCun et al. argue that instead of using the magnitude of weights as a ranking criterion, it is better to minimize the change in the cost function — Equation 2 in our case — after removing one feature at a time (LeCun, Denker, & Solla 1990). More precisely, they expand the cost function in Taylor series. At a local optimum, the first term can be neglected, so those features that minimize the second derivative are eliminated:

$$DJ(i) = \frac{1}{2}\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2, \quad (4)$$

where $DJ(i)$ is the cost function, and w_i is the weight corresponding to the i^{th} feature. Guyon et al. prove that for linear discriminant functions whose cost function J is a quadratic function of w_i, the two criteria are equivalent (Guyon *et al.* 2002). However, for logistic regression, applying the above formula yields to:

$$\frac{\partial^2 l(w)}{\partial w \partial w^T} = -\sum_{i=1}^{N} x_i x_i^T p(x_i; w)(1 - p(x_i; w)), \quad (5)$$

where $p(x_i; w) = Pr(y = 1|x, w)$.

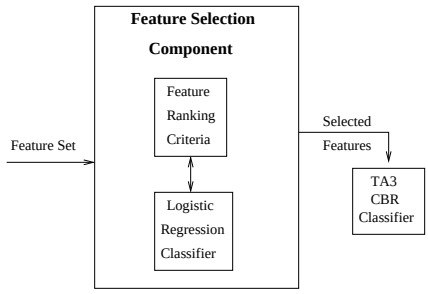

Figure 1: The LR4CBR method

As we demonstrate in the Experimental Results Section, the subset of features with the highest magnitude of weights selected by logistic regression achieves higher accuracy than the subset selected by Fisher and t-test. Figure 1 depicts our proposed method. It first selects a subset of features using $\mid w \mid$ as a feature ranking criterion. The values for $\mid w \mid$ are calculated by training the logistic regression classifier on the training set. Features with the highest ranking magnitude of weights are selected. Our $TA3$ classifier uses the selected feature set to classify the test set.

Prior to applying the above criteria, the data sets are normalized such that all features (regressor variables) have the same mean and the same variance. Since there are thousands of features in the microarray and mass spectrometry data sets, features are eliminated in chunks; the classifier is trained only once, and 90% of "non-informative" features are removed. However, better results might be obtained by removing one feature at a time, and training the classifier on the remaining features before removing the next feature.

The $TA3$ Case-based Reasoning System

Our method is applicable to any CBR system; however, 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) (Jurisica *et al.* 1998) and protein crystal growth (Jurisica *et al.* 2001). This section briefly describes the system.

Case Representation in $TA3$

A case C corresponds to a real world situation, represented as a finite set of attribute/value pairs (Jurisica *et al.* 1998). There are two types of cases: (1) an input case (query) that describes the problem and is represented as a case without a solution; and (2) a retrieved case, which is a case stored in a case-base that contains both a problem description and a solution.

In classification tasks, each case has at least two components: problem description and a 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, which enables treating each partition separately during case retrieval.

Case Retrieval in $TA3$

The retrieval component is based on a modified nearest-neighbor matching (Wettschereck & Dietterich 1995). 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 (Jurisica, Glasgow, & Mylopoulos 2000).

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. Formally, a context is defined as a finite set of attributes with associated constraints on their values:

$$\Omega = \{<a_0 : CV_0>, ..., <a_k : CV_k>\},$$

where a_i is an attribute name and the constraint CV_i specifies the set of "allowable" values for attribute a_i. By selecting only certain features for matching and imposing constraints on feature values, a context controls what is and what is not 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.

Case Adaptation in $TA3$

The adaptation process in CBR manipulates the solution of the retrieved case to better fit the query. Let $x_1, ..., x_k$ denote the k cases retrieved from the case-base that are nearest to query x_q. The method called distance-weighted nearest-neighbor decides on the class label of the query by applying the following formula (Mitchell 1997):

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

where

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

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

Experimental Results

In this section, we compare the FS methods discussed above on two microarray data sets. We then evaluate our proposed method on two publicly available mass spectrometry data sets. Finally, we discuss the list of cancer-specific and control-specific biomarkers selected by the LR4CBR method and markers identified by other studies.

Microarray Data Sets

The experiments have been performed on the following publicly available microarrays data sets:

1. **Leukemia data set**: The data set contains data of 72 leukemia patients, with 7,129 expression levels for each sample [2] (Golub *et al.* 1999). 47 samples belong to type I Leukemia (called Acute Lymphoblastic Leukemia) and 25 samples belong to type II Leukemia (called Acute Myeloid Leukemia).
2. **Lung data set**: The data set taken from the Ontario Cancer Institute [3] contains 39 samples with 18,117 expression levels for each sample. Samples are pre-classified into "recurrence" (23 samples) and "non-recurrence" (16 samples). Missing values were imputed using the KNNimpute software, which is based on the weighted k-nearest-neighbor method (Troyanskaya *et al.* 2001).

Mass Spectrometry Data Sets

The two mass spectrometry data sets (Sorace & Zhan 2003; Zhu *et al.* 2003) discussed in this paper are both provided online at the National Institutes of Health and Food and Drug administration Clinical Proteomics Program Databank. [4]

1. **Ovarian data set 8-7-02**: The ovarian data set 8-7-02 consists of 162 MS spectra from ovarian cancer patients and 91 individuals without cancer (control group) with 15,154 mass-to-charge ratios (m/z values) for each serum.
2. **Ovarian data set 4-3-02**: The 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 of varying intensities.

Feature Selection Results

We used accuracy and classification error to compare several FS methods. Since certain cases cannot be uniquely labeled, the $TA3$ classifier categorizes them as "undecided". In Table 1, *logistic regression I* refers to the first method explained in the LR4CBR Section (selecting the highest ranking magnitude of weights), and *logistic regression II* refers to the FS criterion proposed by LeCun et al. (LeCun, Denker, & Solla 1990).

In order to compare various FS methods for the Leukemia data set, we used the training and the test sets suggested by the data set provider, i.e., 38 samples in the training set, and 34 samples in the test set. As Table 1 shows the accuracy of $TA3_{Leukemia}$ [5] improves using the methods described in the previous section, compared to the case where no FS method is applied.

Leave-one-out cross-validation (LOOCV) was used to compare the various FS methods on the lung data set. In this method, the $TA3_{Lung}$ classifier is successively trained on $n-1$ data points and tested on the remaining one. The results are averaged over 20 trials.

As Table 1 shows, logistic regression outperforms the other FS methods, when the magnitude of weights is used as a ranking criterion. All methods selected the 10% highest ranked features from the whole feature set.

Table 1: Accuracy of $TA3_{Leukemia}$ and $TA3_{Lung}$ with various feature selection methods

Leukemia Data Set			
Method	Accuracy	Error	Undecided
Fisher	74%	23%	3%
T-Test	71%	29%	0
Logistic Regression I	76%	24%	0
Logistic Regression II	71%	29%	0
Lung Data Set			
Method	Accuracy	Error	Undecided
Fisher	55%	25%	20%
T-Test	55%	30%	15%
Logistic Regression I	65%	25%	10%
Logistic Regression II	60%	30%	10%

LR4CBR Results

We used 10-fold cross-validation to evaluate our method, as Table 2 shows, the $TA3$ classifier was 90% accurate when the entire feature set was used, and the LR4CBR method improved the prediction accuracy of $TA3_{8-7-02}$ to 98%. As the Table shows, the classification error reduced from 9.2% to 2%. Recall that cases in which there is a tie, so that $TA3$ cannot decide on its class label are labeled "undecided".

These two ovarian data sets have been previously analyzed (Sorace & Zhan 2003; Zhu *et al.* 2003). Sorace et al. report 100% specificity and 100% sensitivity when they split the ovarian data set 8-7-02 randomly into 125 training and 128 test set. They use the two-sided Wilcoxon test to compare the intensity between controls and cancers in the training set, and then they apply three rules to select a subset of biomarkers (Sorace & Zhan 2003). Although their results are impressive, the rules are extracted in an "ad hoc" way, and might not be applicable to other similar data sets.

Zhu et al. split the 4-3-02 data set randomly into 100 training samples and 116 test samples. After they select a subset of 18 biomarkers using t-test, they apply k-nearest-neighbor (k=5) to classify the test set. They report 100% specificity and 100% sensitivity (Zhu *et al.* 2003). Since we used 10-fold cross-validation, our results are not comparable with them, though we tested $TA3$ with their 18 selected biomarkers. It appeared that the classifier is not as accurate as in the case where the LR4CBR method is applied (third and seventh rows of Table 2). As is typically found in most studies of feature selection, depending on the induction bias of the classifier, different subsets of biomarkers distinguish between cancer and normal samples. Further valida-

[2] `http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_menu.cgi`

[3] `http://www.cs.toronto.edu/~juris/publications/data/CR02Data.txt`

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

[5] $TA3_X$ denotes application of $TA3$ into a domain X.

tion, which is beyond the scope of this paper, will be able to determine which list of biomarkers is biologically more "informative" for diagnosis or treatment prediction for ovarian cancer patients.

When the whole feature set is used, the accuracy is 79.2% (see Table 2). LR4CBR improves the prediction accuracy of $TA3_{4-3-02}$ from 79.2% to 95.4%, while the 18 biomarkers selected by Zhu et al. (Zhu *et al.* 2003) improves the accuracy from 79.2% to 86%. Table 3 displays the list of 15 biomarkers selected by the LR4CBR method. In data set 8-7-02, according to Baggerly et al. (Baggerly, Morris, & Coombes 2004), biomarker 435.46 is the most useful among the seven features reported by (Petricoin *et al.* 2002), Our algorithm selects that biomarker as well (Table 3). Also, it is worth mentioning that Baggerly et al. (Baggerly, Morris, & Coombes 2004) report biomarker 244.95 as the best single biomaker for data set 8-7-02, and this biomarker also appears in our list. It is also interesting to see that biomarker 434.69 is in the list of selected biomarkers reported by Zhu et al. (Zhu *et al.* 2003).

Table 2: Accuracy of $TA3_{8-7-02}$ and $TA3_{4-3-02}$

Ovarian Data Set 8-7-02			
Method	Accuracy	Error	Undecided
TA3(no FS)	90%	9.2%	0.8%
18 Biomarkers	92.5%	6.3%	1.2%
LR4CBR	98%	2%	0%
Ovarian Data Set 4-3-02			
Method	Accuracy	Error	Undecided
TA3(no FS)	79.2%	18.5%	2.3%
18 Biomarkers	86%	8%	6%
LR4CBR	95.4%	4.6%	0%

Table 3: Biomarkers selected by LR4CBR for the ovarian data set 8-7-02 and 4-3-02

Ovarian Data Set 8-7-02				
Markers(m/z)				
244.66	**244.95**	245.24	245.54	245.83
246.12	261.58	261.89	417.35	417.73
434.69	435.07	**435.46**	435.85	436.24
Ovarian Data Set 4-3-02				
Markers(m/z)				
0.0386	0.0463	0.0504	0.0735	0.0786
0.0895	0.1468	0.2451	0.4033	0.4153
0.4274	0.5170	0.5306	0.6445	1518.8719

Conclusions and Future Work

Life sciences domains, such as gene and protein expression profiling, are natural applications for CBR systems, since CBR systems can perform remarkably well on complex and poorly formalized domains. However, due to the large number of attributes in each case, CBR classifiers, similarly to other learning systems, suffer from the "curse of dimensionality". Integrating CBR systems with feature selection techniques improves the prediction accuracy of CBR classifiers by removing "non-informative" features in each group.

In this paper, we have proposed using logistic regression as a filter feature selection method for the $TA3$ CBR classifier. According to our experiments on microarray data sets, logistic regression performs more accurately than Fisher and t-test. We also evaluated our method on two public mass spectrometry data sets, and showed that LR4CBR improves the accuracy from 90% to 98% on the ovarian data set 8-7-02 and from 79.2% to 95.4% on the ovarian data set 4-3-02.

Future investigation may further exploit the advantage of Telos-style categories in $TA3$ for classification tasks, and validate the system on diverse high-dimensional data sets. We plan to evaluate the classifier with wrapper as well as *hybrid* feature selection techniques: a combination of filter and wrapper approach (Das 2001). Because of the high dimensionality of our data sets, we removed features in "larger" chunks, though a method that removes features in "smaller" chunks may lead to better performance. Also, we may consider feature redundancy as well as feature relevance, as Yu and Liu (Yu & Liu 2004) discuss that feature relevance alone is insufficient for efficient feature selection of high-dimensional data.

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

Aha, D. W., and Bankert, R. 1994. 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*, 106–112. Menlo Park, CA: AAAI Press.

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

Baggerly, K.; Morris, J.; and Coombes, K. 2004. Reproducibility of seldi-top protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 20(5):777–785.

Das, S. 2001. Filters, wrappers and a boosting based hybrid for feature selection. In Brodley, C. E., and Danyluk, A. P., eds., *Proceedings of the Eighteenth International Conference on Machine Learning*, 74–81. Williamstown, MA, USA: Morgan Kauffmann.

Devore, J. 1995. *Probability and statistics for engineering and the sciences*. Duxbury Press.

Golub, T.; Slonim, D.; Tamayo, P.; Huard, C.; Gassenbeek, M.; Mesirov, J.; Coller, H.; Loh, M.; Downing, J.; Caligiuri, M.; Bloomfield, C.; and Lander, E. 1999. Molecular classification of cancer: class discovery and class pre-

diction by gene expression monitoring. *science* 286:531–537.

Guyon, I.; Weston, J.; Barnhill, S.; and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1/3):389–422.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2001. *The elements of statistical learning*. Springer.

Jaeger, J.; Sengupta, B.; and Ruzzo, W. 2003. Improved gene selection for classification of microarrays. In *Pacific Symposium on Biocomputing*, 8:53–64.

Jurisica, I., and Glasgow, J. 2004. Application of case-based reasoning in molecular biology. *Artificial Intelligence Magazine, Special issue on Bioinformatics* 25(1):85–95.

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

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

Jurisica, I.; Glasgow, J.; and Mylopoulos, J. 2000. Incremental iterative retrieval and browsing for efficient conversational CBR systems. *International Journal of Applied Intelligence* 12(3):251–268.

Kohavi, R., and John, G. H. 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97(1-2):273–324.

Leake, D. B., ed. 1996. *Case-based reasoning: experiences, lessons, and future directions*. AAAI Press/MIT Press.

LeCun, Y.; Denker, J.; and Solla, S. 1990. Optimal brain damage. In Touretzky, D., ed., *Advances in Neural Information Processing Systems*, volume 2, 598–605. Morgan Kaufmann, San Mateo, CA.

Lenz, M.; Bartsch-Sporl, B.; Burkanrd, H.; and Wess, S., eds. 1998. *Case-Based Reasoning: experiences, lessons, and future directions*. Springer.

Mitchell, T. M. 1997. *Machine Learning*. McGraw-Hill.

Mukherjee, S. 2003. Classifying microarray data using support vector machines. In Berrar, D.; Dubitzky, W.; and Granzow, M., eds., *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers. chapter 9, 166–185.

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.; and Liotta, L. A. 2002. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359:572–577.

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

Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; and Altman, R. B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17:520–525.

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

Wu, B.; Abbott, T.; Fishman, D.; McMurray, W.; Mor, G.; Stone, K.; Ward, D.; Williams, K.; and Zhao, H. 2003. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19(13):1636–1643.

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

Yu, L., and Liu, H. 2004. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research* 5:1205–1224.

Zhu, W.; Wang, X.; Ma, Y.; Rao, M.; Glimm, J.; and Kovach, J. S. 2003. 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):14666–14671.