Diagnosis of Ovarian Cancer

Based on Mass Spectrum of Blood Samples

by

Hong Tang

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Co-Major Professor: Eugene Fink, Ph.D.
Co-Major Professor: Lihua Li, Ph.D.
Member: Dmitry B. Goldgof, Ph.D.

Date of Approval:
July 22, 2003

Keywords: Data Mining, Medical Application, Decision Trees,
Support Vector Machines, Neural Networks.

Acknowledgements

I am grateful to my faculty supervisors, Eugene Fink and Lihua Li, for their help and support. I thank Dmitry Goldgof for his valuable comments and suggestions. I also appreciate the help of my fellow graduate students: Yong Chu, Jianli Gong, Yelena Mukomel and Savvas Nikiforou.

I am deeply indebted to my husband, Peng Zhang, and my daughter, Alice Zhang, for their support, encouragement and patience throughout my studies. I am also grateful to my parents, Fahui Tang and Cuiwen Song, and my brother, Ming Tang.

Table of Contents

List of Tables

List of Figures

Diagnosis of Ovarian Cancer
Based on Mass Spectrum of Blood Samples

Hong Tang

ABSTRACT

The early detection of cancer is crucial for successful treatment, and medical researchers have investigated a number of early-diagnosis techniques. Recently, they have discovered that some cancers affect the concentration of certain molecules in the blood, which allows early diagnosis by analyzing the blood mass spectrum. Researchers have developed several techniques for the analysis of the mass-spectrum curve, and used them for the detection of prostate, ovarian, breast, bladder, pancreatic, kidney, liver and colon cancers.

We have continued this work and applied data mining to the diagnosis of ovarian cancer based on the mass-spectrum curve. We have identified the most informative points of this curve, and then used decision trees, support vector machines, and neural networks to determine the differences between the curves of cancer patients and healthy people.

Chapter 1 Introduction

The development of tools for the early cancer diagnosis is a major open problem, and clinicians have investigated a variety of diagnosis techniques. Recently, they have discovered that cancer may affect the blood mass spectrum, and studied diagnosis methods based on the analysis of mass-spectrum data, which provide information about proteins and their fragments [Bakhtiar and Tse, 2000; Yates, 2000; Bakhtiar and Nelson, 2001].

Researchers use two main techniques for generating mass spectra, which are called "matrix-assisted laser desorption and ionization" [Valerio *et al.*, 2001; Wu *et al.*, 2003] and "surface-enhanced laser desorption and ionization" [Adam *et al.*, 2001; Wulfkuhle *et al.*, 2001; Chapman, 2002; Issaq *et al.*, 2002; Wellmann *et al.*, 2002]. The resulting mass spectrum is a curve (Figure 1.1), where the x-axis shows the ratio of the weight of a specific molecule to its electrical charge, and the y-axis is the signal intensity for the same molecule. The mass-spectrum analysis is a fast inexpensive procedure based on a sample of a patient's blood, and it may potentially allow cancer screening with little discomfort to a patient.

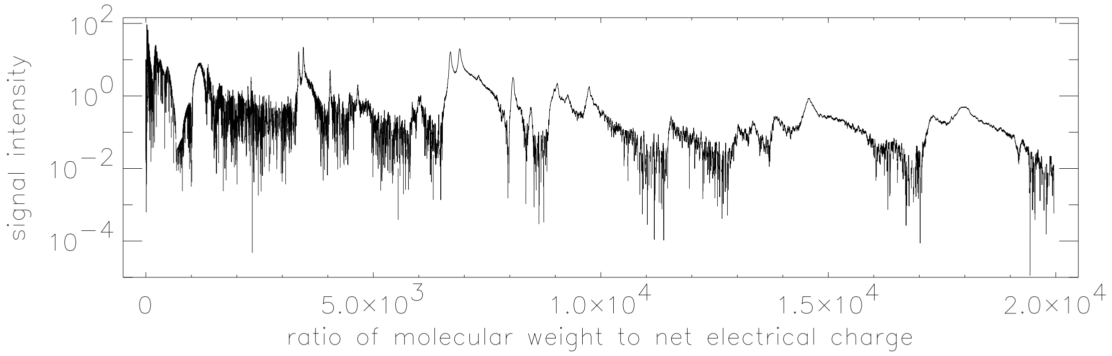


Figure 1.1: Mass-Spectrum Curve.

Table 1.1: Data Sets Used in the Reported Work.

Data set	Number of cases	
	Cancer	Healthy
1	100	116
2	100	116
3	162	91

Medical researchers have developed several techniques for analyzing the mass-spectrum data, which allow the diagnosis of various cancers, including ovarian, breast, prostate, bladder, pancreatic, kidney, liver and colon cancers. The effectiveness of these techniques varies across cancer types, methods for generating mass spectra, and algorithms for analyzing the resulting data. Clinicians use three standard measures of the effectiveness of diagnosis techniques: sensitivity, specificity and accuracy. The *sensitivity* is the probability of the correct diagnosis for a patient with cancer, the *specificity* is the chances of the correct diagnosis for a healthy person, and the *accuracy* is the chances of the correct diagnosis for the overall population of healthy and sick people. The sensitivity of the mass-spectrum diagnosis techniques has varied from 64% to 99%, the specificity has been between 66% and 98%, and the overall accuracy has been between 73% and 98%.

We have continued this work and investigated techniques for the diagnosis of early-stage ovarian cancer. Specifically, we have applied decision-tree learning, support vector machines, and neural networks to identify the differences between the mass spectra of ovarian-cancer patients and those of healthy people. We have used three data sets (Table 1), available at *http://clinicalproteomics.steem.com*. Sets 1 and 2 include the mass spectra of 100 cancer patients and 116 healthy people, whereas Set 3 includes the data of 162 cancer patients and 91 healthy people. Each mass-spectrum curve consists of 15,155 points.

The experiments have confirmed that the mass spectra allow the diagnosis of ovarian cancer. The sensitivity of the developed technique varies from 85% to 99%, depending on the data set, its specificity is between 81% and 99%, and its accuracy is between 82% and 99%.

Chapter 2 Previous Work

Medical researchers have developed techniques for the detection of early cancer based on *protein markers,* which are certain molecules in body tissues and fluids [Poon and Johnson, 2001], but these techniques are often inaccurate. For example, the specificity of an antigen method for the prostate-cancer detection is only 25–30%, although its sensitivity is high [Adam *et al.*, 2001]; as another example, the sensitivity of a similar method for breast cancer is 23%, and its specificity is 69% [Li *et al.*, 2002]. Recently, researchers have developed a new cancer-detection method, based on the application of data mining to the mass spectra of patients' tissue cells, blood, serum and other body fluids [Alaiya *et al.*, 2000; Banks *et al.*, 2000; Celis *et al.*, 2000; Chambers *et al.*, 2000; Paweletz *et al.*, 2000; Adam *et al.*, 2001; Poon and Johnson, 2001; Srinivas *et al.*, 2001; Vlahou *et al.*, 2001; Wulfkuhle *et al.*, 2001; Fung and Enderwick, 2002; Issaq *et al.*, 2002; Petricoin *et al.*, 2002a; Petricoin *et al.*, 2002c; Petricoin and Liotta, 2002; Wulfkuhle *et al.*, 2003].

2.1 Peaks in Mass Spectra

Some researchers have analyzed mass-spectrum curves using the Ciphergen System software, which helps to identify major peaks. Hlavaty *et al.* [2001] found that a 50.8k Dalton protein peak was present in all prostate-cancer samples, and absent in all samples of healthy people. Watkins *et al.* [2001] used the same method to detect breast, colon and prostate cancers. They correctly identified 100% of breast cancer cases and ruled out 96% of noncancer cases. For colon cancer, they correctly identified 100% of cancer cases and ruled out 86% of noncancer cases. For prostate cancer, their results were 100% correct for both cancer and noncancer cases. Sauter *et al.* [2002] analyzed mass-spectrum curves of the nipple aspirate fluid over the 5–40k Dalton range, and identified five relevant peaks. The most relevant peaks were 6.5k Dalton and 15.9k Dalton, and their use gave 84% sensitivity and 100% specificity.

2.2 Decision Trees

Adam *et al.* [2002] applied decision-tree learning to the blood mass spectra of prostate-cancer patients. They used the Ciphergen System software for peak detection, and decision trees for classification based on the intensity of nine highest peaks, which gave 96% accuracy, 83% sensitivity and 97% specificity. They also experimented with biostatistical algorithms, genetic clustering and support vector machines, which gave accuracy between 83% and 90%. Qu *et al.* [2002] applied a boosted decision tree method, using the same data and features as Adam *et al.* [2002]. They developed two new classifiers, called AdaBoost and Boosted Decision Stump Feature Selection. For AdaBoost, the sensitivity was 98.5% with the 95% confidence interval of 96.5–99.7%, and the specificity was 97.9% with the 95% confidence interval of 95.5–99.4%. For Boosted Decision Stump Feature Selection, the sensitivity was 91.1% with the 95% confidence interval of 86.9–94.6%, and the specificity was 94.3% with the 95% confidence interval of 90.7–97.1%.

2.3 Neural Networks

Ball *et al.* [2002] applied back-propagation neural networks to determine astroglial tumor grade (1 or 2), which gave 100% accuracy. Poon *et al.* [2003] used neural networks to distinguish hepatocellular carcinoma from chronic liver disease, which gave 92% sensitivity and 90% specificity.

2.4 Clustering

Petricoin *et al.* [2002a] combined a genetic algorithm with self-organizing cluster analysis for identifying ovarian cancer. The sensitivity of their technique was 100%, with the 95% confidence interval of 93–100%, and the specificity was 95%, with the 95% confidence interval of 87–99%. They also applied their technique to diagnose prostate cancer [Petricoin *et al.*, 2002b], which gave 95% sensitivity with the 95% confidence interval of 82–99%, and 78% specificity with the 95% confidence interval of 72–83%.

Poon *et al.* [2003] applied two-way hierarchical clustering to distinguish hepatocellular carcinoma from chronic liver disease; however, they did not report its sensitivity, specificity or accuracy.

2.5 Other Methods

Valerio *et al.* [2001] applied the statistical χ^2 test to the mass spectra of thirteen pancreatic cancer patients, nine chronic pancreatitis patients and ten healthy people, and found unique protein peaks for each of the three groups; however, they did not report the sensitivity, specificity or accuracy of their method. Cazares *et al.* [2002] analyzed mass spectra of prostate cancer; they used the Ciphergen System software for peak detection, and logistic regression for classification, which gave 93% sensitivity and 94% specificity. Wu *et al.* [2003] compared several methods for classification of ovarian cancer, including linear discriminant analysis, quadratic discriminant analysis, nearest neighbors, bagging classification trees, boosting classification trees, support vector machines and random forests; they concluded that the random-forest classification was the most effective.

Chapter 3 New Results

We describe a technique for selecting relevant points of the mass-spectrum curve, and then give results of detecting ovarian cancer based on the values of these points.

3.1 Feature Selection

We view each point of a mass-spectrum curve as a feature, and the corresponding signal intensity as its value. To select relevant features, we calculate the mean intensity values for each point in the mass spectra of the cancer and non-cancer groups, μ_1 and μ_2, and the corresponding standard deviations, σ_1 and σ_2. The mean difference of these intensities is $|\mu_1 - \mu_2|$, and the standard deviation of this difference is $\sqrt{\sigma_1^2 + \sigma_2^2}$. For each point, we determine the ratio of the mean difference to its standard deviation, $|\mu_1 - \mu_2|/\sqrt{\sigma_1^2 + \sigma_2^2}$, and select a given number of points with the greatest ratios.

We impose a lower bound on the distance between selected points, which prevents the selection of points with correlated values. After selecting the point with the greatest ratio, we discard all points within the distance bound from the selected point and choose the second greatest-ratio feature among the remaining points. Then, we discard the points within the distance bound from the second selected point, choose the third greatest-ratio feature among the remaining points, and so on.

3.2 Learning Algorithms

We have experimented with decision trees, support vector machines and neural networks. We have used the C4.5 package (*www.cse.unsw.edu.au/~quinlan*) for learning decision trees [Quinlan, 1993], the SVMFu package (*five-percent-nation.mit.edu/SvmFu*) for constructing support vector machines with linear kernel functions [Burges, 1998; Cristianini and Shawe-Taylor, 2000] and the Cascor 1.2 package (*www.cs.cmu.edu/afs/cs/pr-*

oject/connect/code/supported) for generating neural networks using the cascade-correlation algorithm [Fahlman and Lebiere, 1990; Fausett, 1994; Bishop, 1995]. Cascor starts with a network that has no hidden units, and adds new units one by one, in a two-step process. First, it adds a new hidden unit and connections from the input units and old hidden units to the new unit, and trains the weights of these connections. Second, it adds the connections from the new unit to the output unit and trains their weights.

3.3 Experiments

We have implemented an experimental setup that allows control over the number of features and minimal distance between selected features. We have varied the number of features from 1 to 64, and the minimal distance from 1 to 1024. For each combination of settings, we have used eighteen-fold cross-validation to evaluate the three learning algorithms. In Figures 3.1–3.7, we show the dependency of the accuracy on the control variables. In Table 3.1, we give the minimal and maximal sensitivity, specificity and accuracy for decision trees, support vector machines and neural networks.

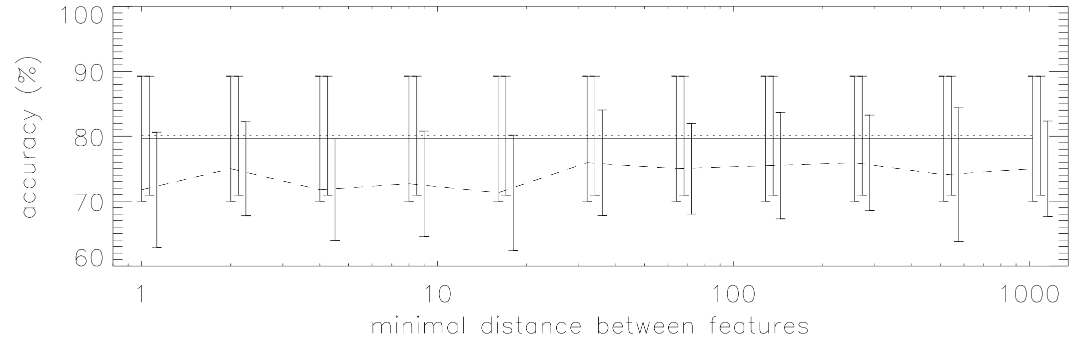
We have determined the number of features and minimal distance between features that lead to the highest accuracy (Table 3.2). The optimal number of features varies from four to thirty-two, depending on the learning technique and data set. We have also constructed the learning curves for the optimal choice of parameters (Figures 3.8–3.10); these curves show the dependency of the accuracy on the training-set size. The results show that all three techniques reach the maximal accuracy after processing about one hundred learning examples.

Table 3.1: Effectiveness of Ovarian-Cancer Diagnosis. We Show the Minimal (Min) and Maximal (Max) Accuracy, Sensitivity and Specificity for Each of the Learning Techniques.

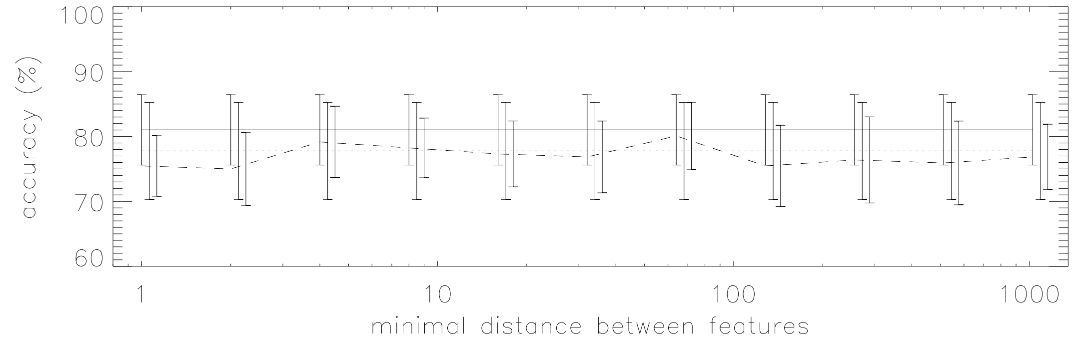
		Decision trees		SVM		Neural nets	
		Min	Max	Min	Max	Min	Max
Data set 1	Accuracy	75%	82%	76%	83%	67%	82%
	Sensitivity	68%	86%	75%	85%	66%	85%
	Specificity	72%	81%	74%	85%	67%	84%
Data set 2	Accuracy	81%	94%	78%	94%	75%	96%
	Sensitivity	81%	95%	70%	98%	74%	94%
	Specificity	77%	96%	79%	94%	76%	97%
Data set 3	Accuracy	96%	99%	88%	99%	94%	99%
	Sensitivity	96%	99%	85%	100%	94%	100%
	Specificity	91%	100%	95%	99%	92%	99%

Table 3.2: Control-Variable Values That Lead to the Maximal Accuracy, and the Corresponding Accuracy, Sensitivity and Specificity.

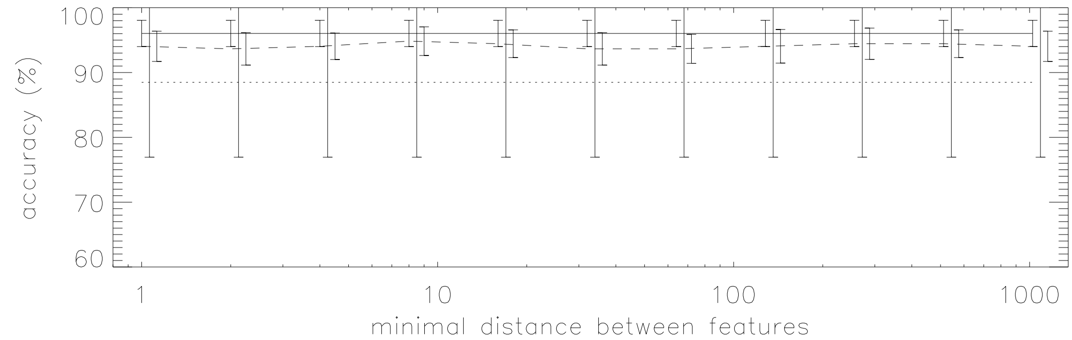
		Num. of features	Minimal distance	Accuracy	Sensitivity	Specificity
Data set 1	Decision trees	4	1	82%	86%	78%
	SVM	32	16	83%	82%	84%
	Neural nets	32	256	82%	80%	84%
Data set 2	Decision trees	8	4	94%	92%	96%
	SVM	4	2	94%	96%	93%
	Neural nets	32	1	96%	93%	98%
Data set 3	Decision trees	8	64	99%	98%	100%
	SVM	16	8	99%	100%	99%
	Neural nets	16	2	99%	100%	99%



(a) Data set 1.

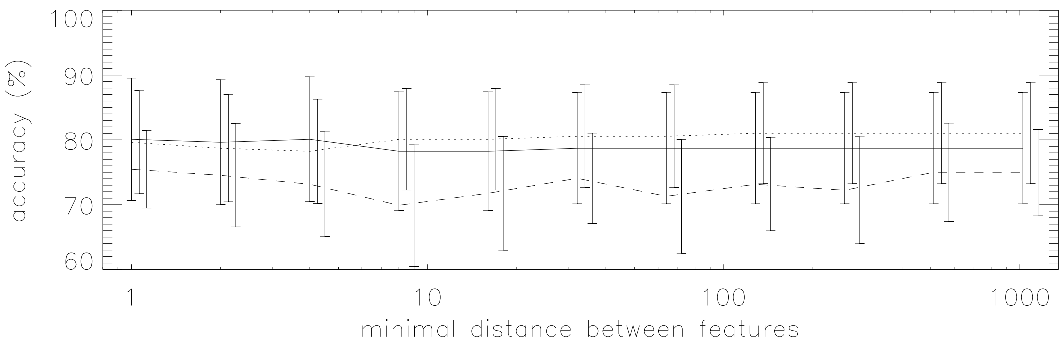


(b) Data set 2.

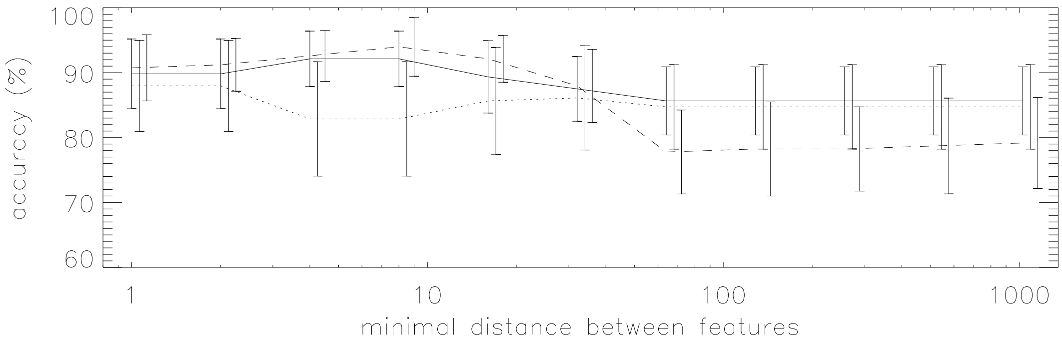


(c) Data set 3.

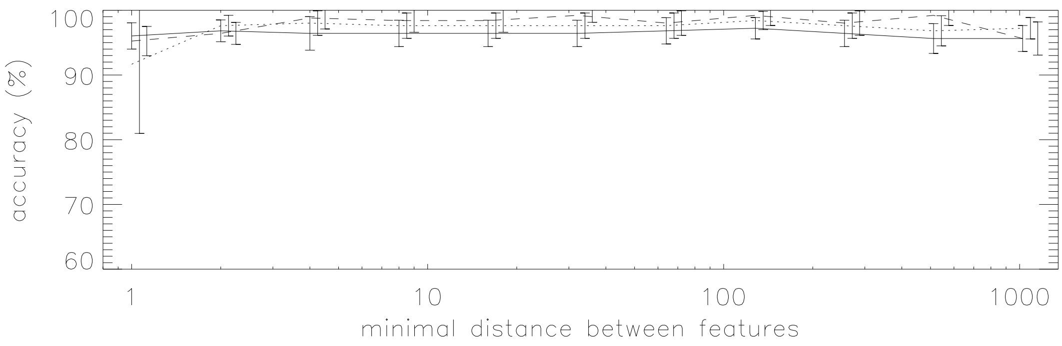
Figure 3.1: Experiments with One Feature. We Plot the Accuracy for Decision Trees (Solid Lines), Support Vector Machines (Dotted Lines) and Neural Networks (Dashed Lines). The Vertical Bars Show the Standard Deviation for Decision Trees (Left), Support Vector Machines (Middle) and Neural Networks (Right).



(a) Data set 1.

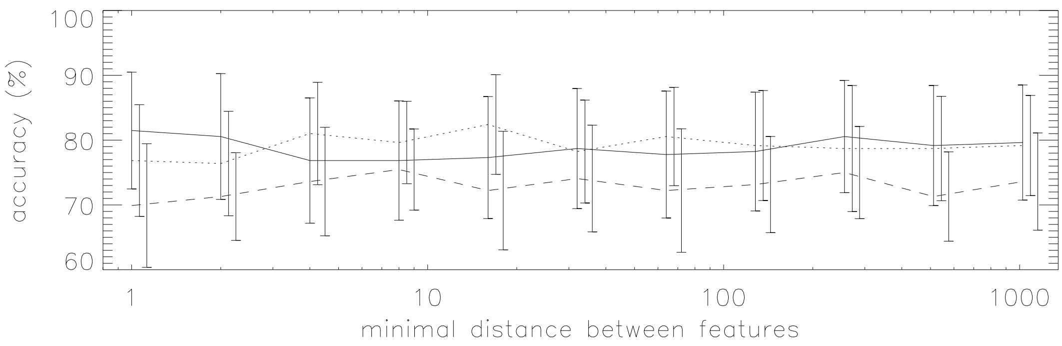


(b) Data set 2.

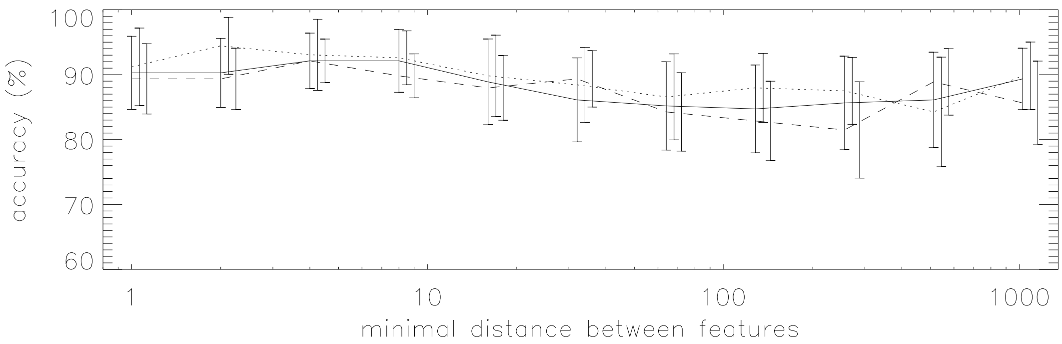


(c) Data set 3.

Figure 3.2: Experiments with Two Features. The Legend Is the Same as in Figure 3.1.



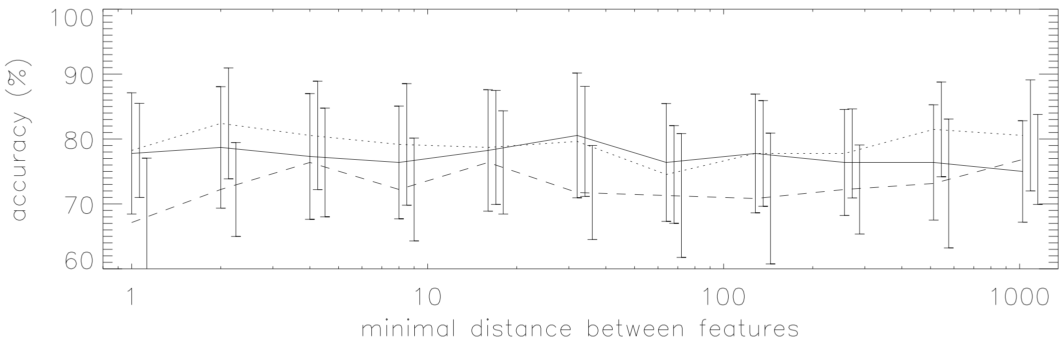
(a) Data set 1.



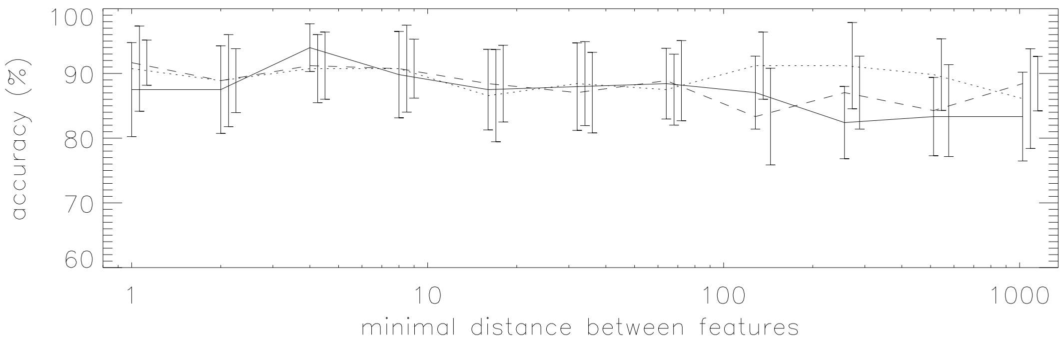
(b) Data set 2.

(c) Data set 3.

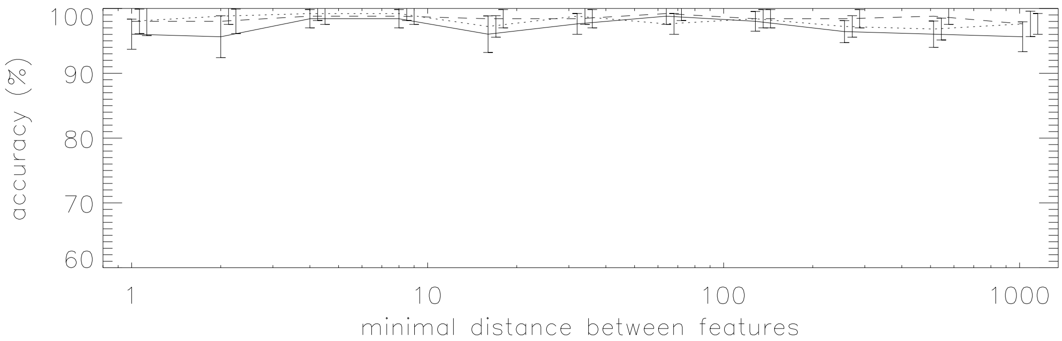
Figure 3.3: Experiments with Four Features. The Legend Is the Same as in Figure 3.1.



(a) Data set 1.

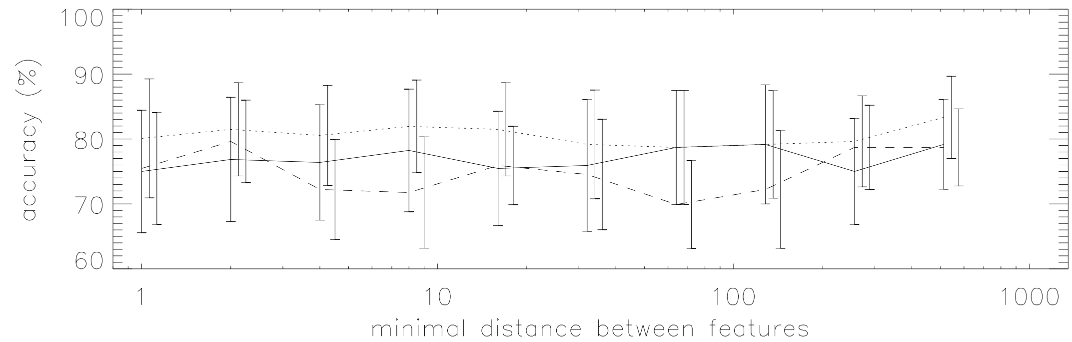


(b) Data set 2.

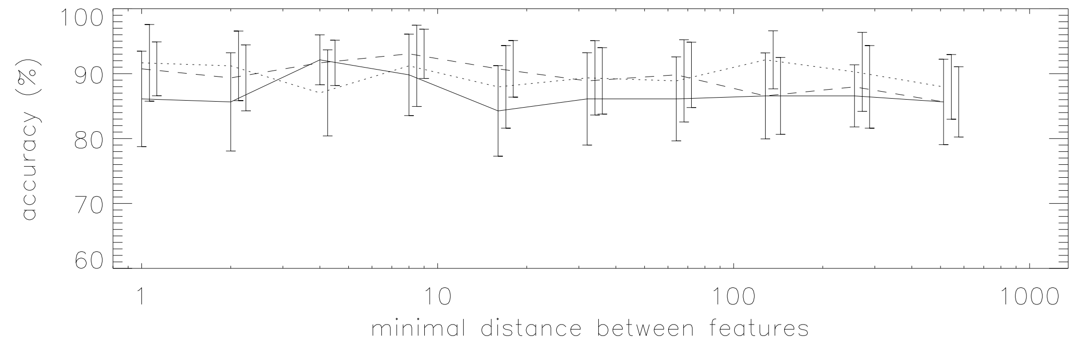


(c) Data set 3.

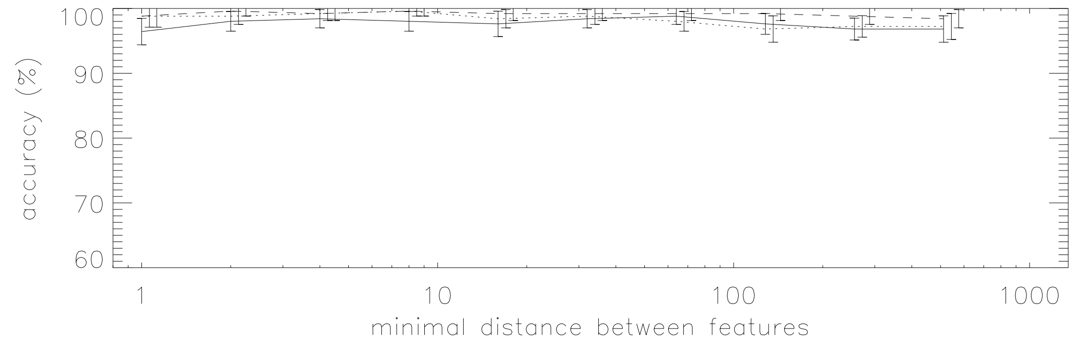
Figure 3.4: Experiments with Eight Features. The Legend Is the Same as in Figure 3.1.



(a) Data set 1.

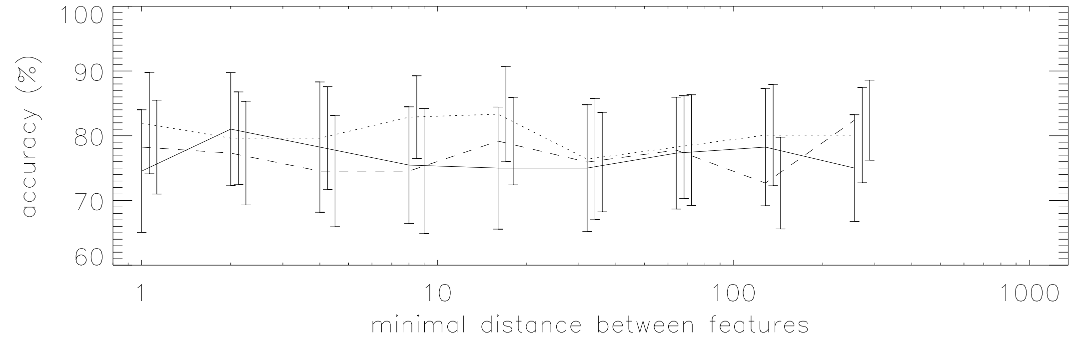


(b) Data set 2.

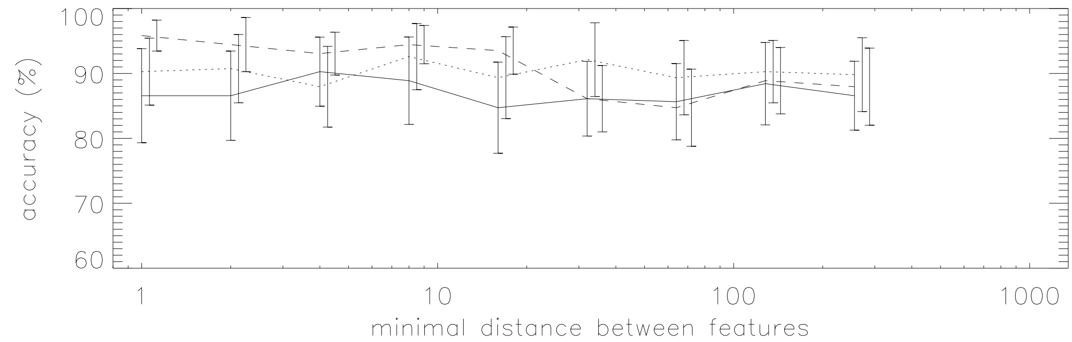


(c) Data set 3.

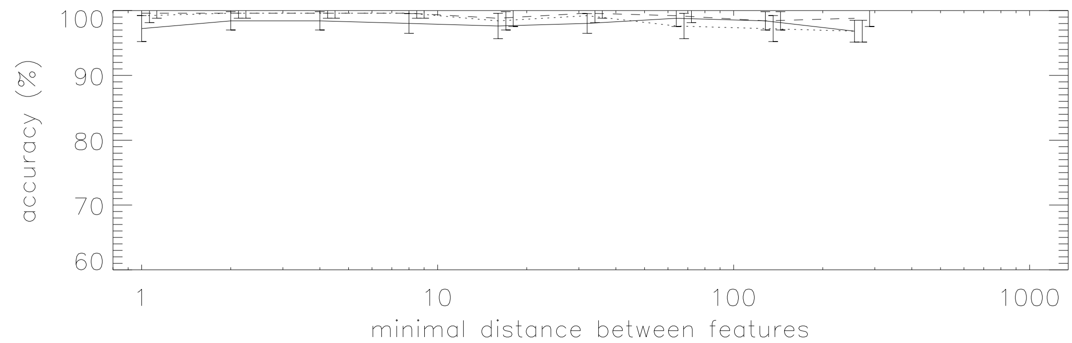
Figure 3.5: Experiments with Sixteen Features. The Legend Is the Same as in Figure 3.1.



(a) Data set 1.

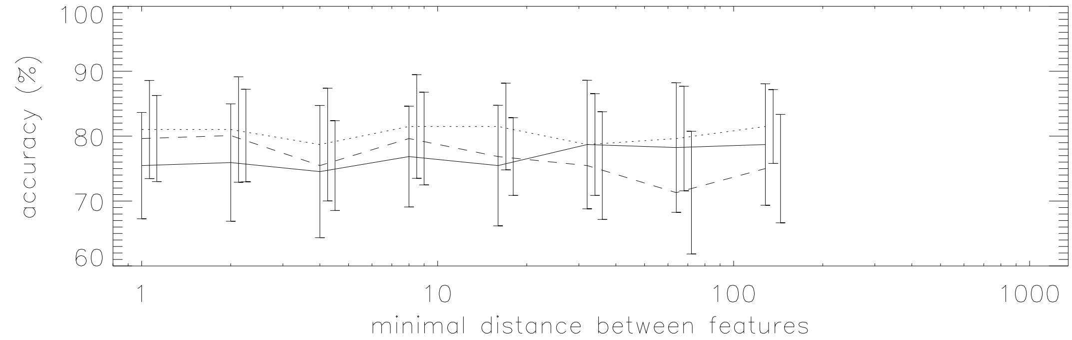


(b) Data set 2.

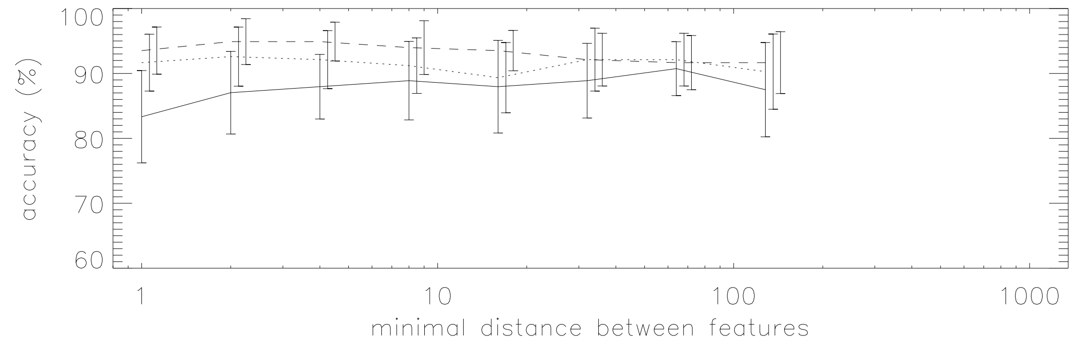


(c) Data set 3.

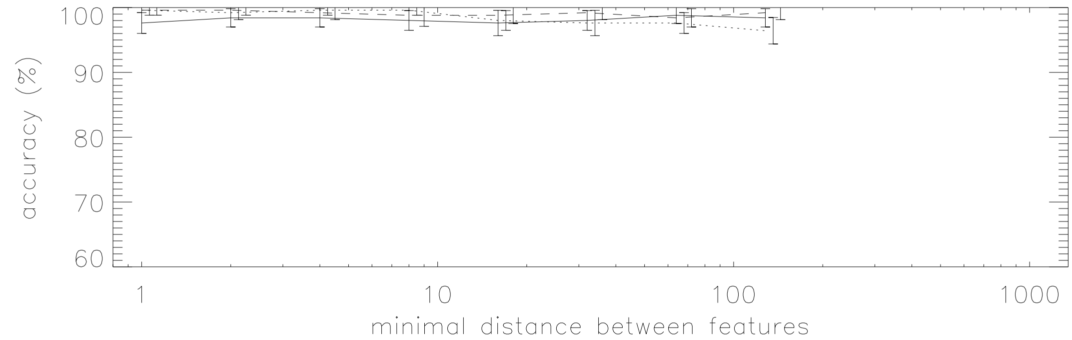
Figure 3.6: Experiments with Thirty-Two Features. The Legend Is the Same as in Figure 3.1.



(a) Data set 1.

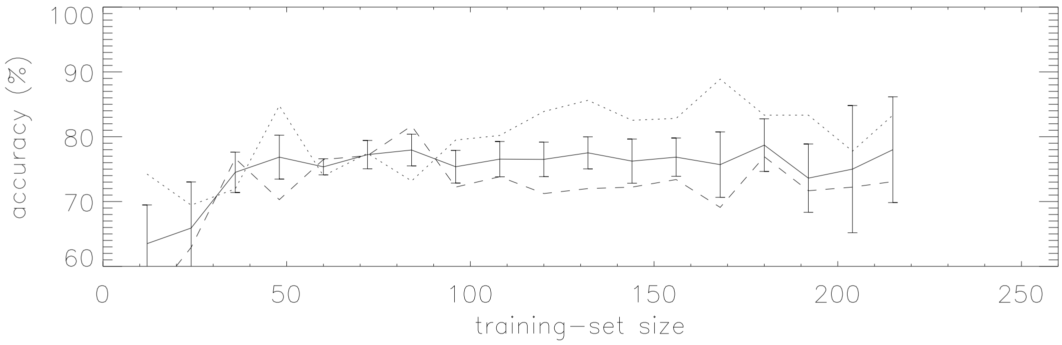


(b) Data set 2.

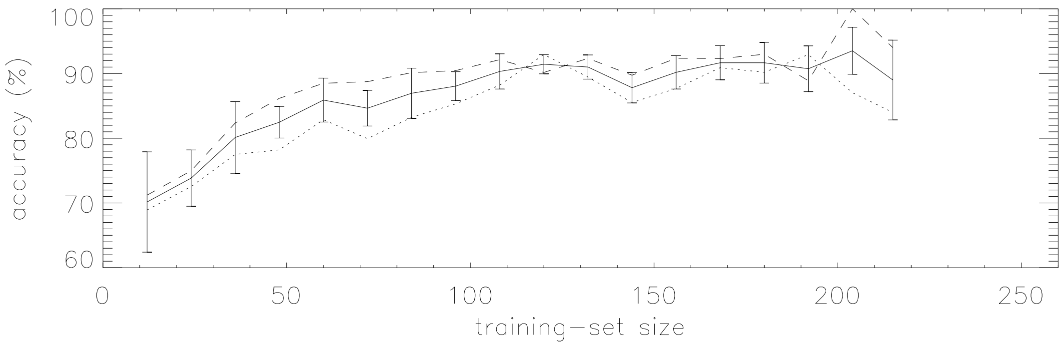


(c) Data set 3.

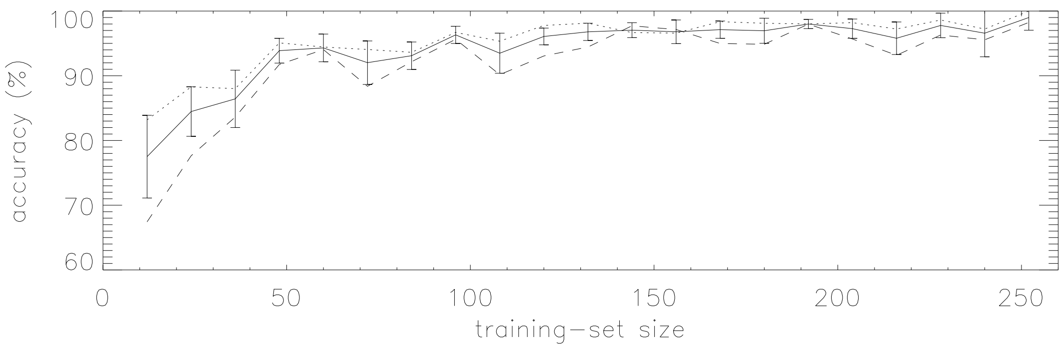
Figure 3.7: Experiments with Sixty-Four Features. The Legend Is the Same as in Figure 3.1.



(a) Data set 1.

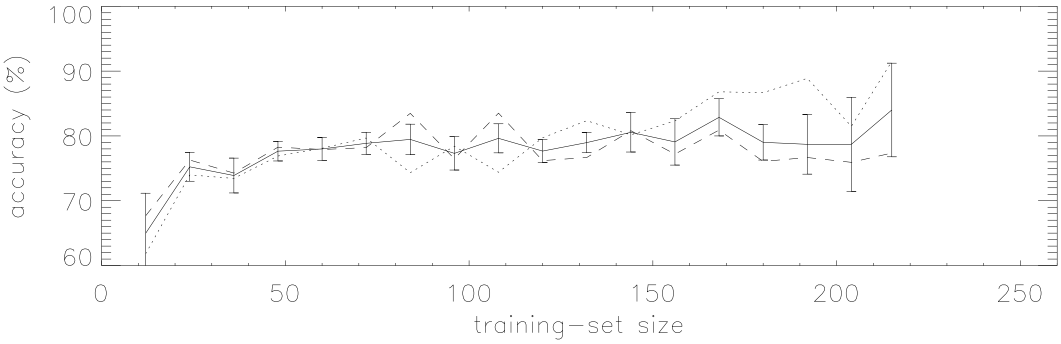


(b) Data set 2.

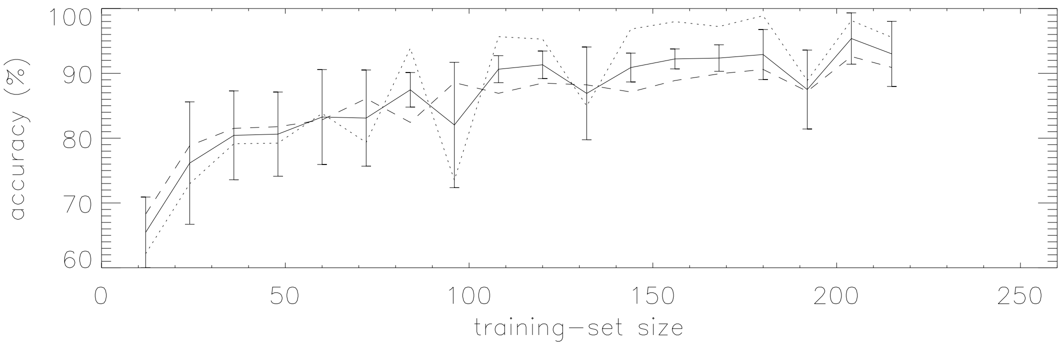


(c) Data set 3.

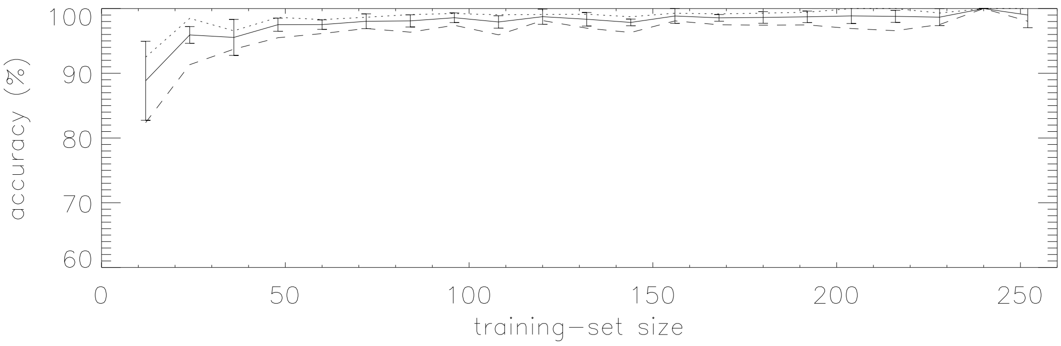
Figure 3.8: Learning Curves for Decision Trees. We Plot the Accuracy (Solid Lines), Sensitivity (Dotted Lines) and Specificity (Dashed Lines). The Vertical Bars Show the Standard Deviation of Accuracy.



(a) Data set 1.

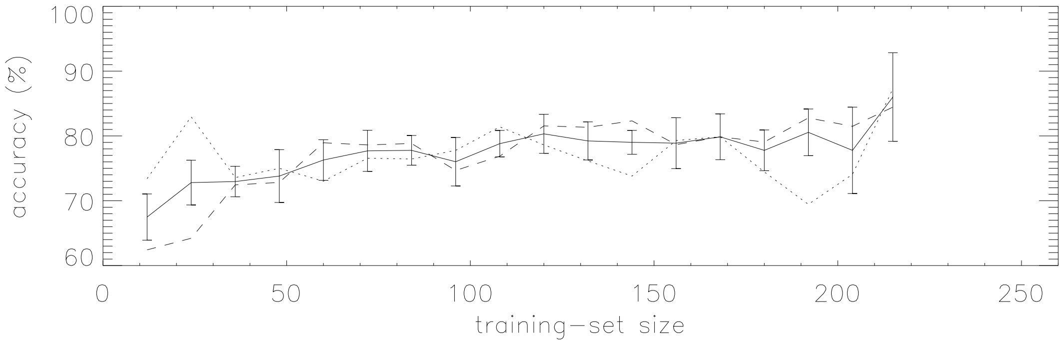


(b) Data set 2.

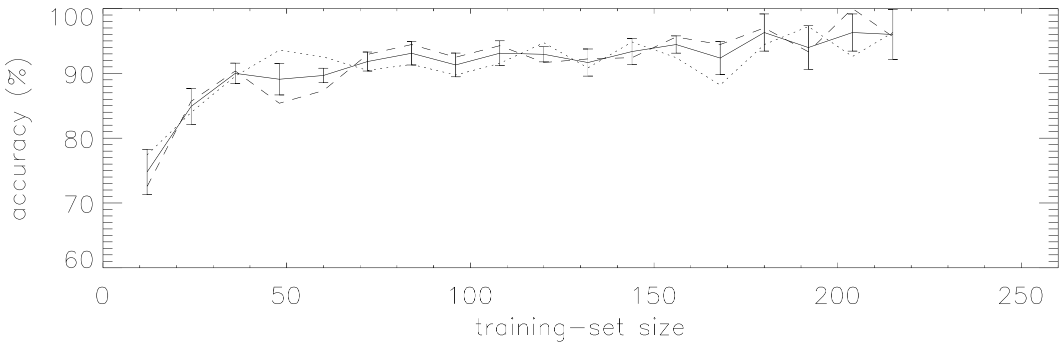


(c) Data set 3.

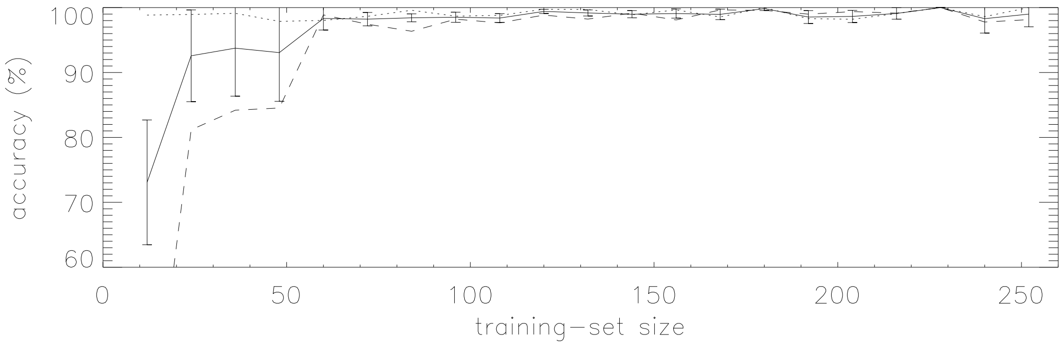
Figure 3.9: Learning Curves for Support Vector Machines. The legend Is the Same as in Figure 3.8.



(a) Data set 1.



(b) Data set 2.



(c) Data set 3.

Figure 3.10: Learning Curves for Neural Networks. The Legend Is the Same as in Figure 3.8.

Chapter 4 Concluding Remarks

We have considered the problem of diagnosing ovarian cancer based on the blood mass-spectrum curve, and identified the relevant points of the curve. We have then applied decision trees, support vector machines, and neural networks to determine the values of these points that indicate ovarian cancer. The effectiveness of these techniques varies across the available data sets; the accuracy of decision trees is between 82% and 99%, the accuracy of support vector machines is between 83% and 99%, and the accuracy of neural networks is between 82% and 99%. The related future work may include experiments with other feature-selection methods, and integration of the developed techniques with genetic algorithms.

References

[Adam *et al.*, 2001] Bao-Ling Adam, Antonia Vlahou, Oliver John Semmes, and George L. Wright. Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics*, 1(10):1264–1270, 2001.

[Adam *et al.*, 2002] Bao-Ling Adam, Yinsheng Qu, John W. Davis, Michael D. Ward, Mary Ann Clements, Lisa H. Cazares, Oliver John Semmes, Paul F. Schellhammer, Yutaka Yasui, Ziding Feng, and George L. Wright. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, 62:3609–3614, 2002.

[Alaiya *et al.*, 2000] Ayodele A. Alaiya, Bo Franzén, Gert Auer, and Stig Linder. Cancer proteomics: From identification of novel markers to creation of artificial learning models for tumor classification. *Electrophoresis*, 21:1210–1217, 2000.

[Bakhtiar and Nelson, 2001] Ray Bakhtiar and Randall W. Nelson. Mass spectrometry of the proteome. *Molecular Pharmacology*, 60(3):405–415, 2001.

[Bakhtiar and Tse, 2000] Ray Bakhtiar and F. L. S. Tse. Biological mass spectrometry: A primer. *Mutagenesis*, 15(5):415–430, 2000.

[Ball *et al.*, 2002] Graham Ball, Saira Mian, F. Holding, R. O. Allibone, J. Lowe, Selman Ali, Gui-Ru Li, S. McCardle, Ian O. Ellis, Colin Creaser, and Robert C. Rees. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*, 18(3):395–404, 2002.

[Banks *et al.*, 2000] Rosamonde E. Banks, Michael J. Dunn, Denis F. Hochstrasser, Jean-Charles Sanchez, Walter Blackstock, Darryl J. Pappin, and Peter J. Selby. Proteomics: New perspectives, new biomedical opportunities. *The Lancet*, 356:1749–1756, 2000.

[Bishop, 1995] Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, UK, 1995.

[Burges, 1998] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[Cazares *et al.*, 2002] Lisa H. Cazares, Bao-Ling Adam, Michael D. Ward, Suhail Nasim, Paul F. Schellhammer, Oliver John Semmes, and George L. Wright. Normal, benign, preneoplastic, and malignant prostate cells have distinct protein expression profiles resolved by surface enhanced laser desorption/ionization mass spectrometry. *Clinical Cancer Research*, 8(8):2541–2552, 2002.

[Celis *et al.*, 2000] Julio E. Celis, Mogens Kruhøffer, Irina Gromova, Casper Frederiksen, Morten Østergaard, Thomas Thykjaer, Pavel Gromov, Jinsheng Yu, Hildur Pálsdóttir, Nils Magnusson, and Torben F. Ørntoft. Gene expression profiling: Monitoring transcription and translation products using DNA microarrays and proteomics. FEBS *Letters*, 480(1):2–16, 2000.

[Chambers *et al.*, 2000] George Chambers, Laura Lawrie, Phil Cash, and Graeme I. Murray. Proteomics: A new approach to the study of disease. *Journal of Pathology*, 192:280–288, 2000.

[Chapman, 2002] K. Chapman. The ProteinChip Biomarker System from Ciphergen Biosystems: A novel proteomics platform for rapid biomarker discovery and validation. *Biochemical Society Transactions*, 30(2):82–87, 2002.

[Cristianini and Shawe-Taylor, 2000] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machine and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK, 2000.

[Fahlman and Lebiere, 1990] Scott E. Fahlman and Christian Lebiere. The cascade-correlation learning architecture. In David S. Touresky, editor, *Advances in Neural Information Proceeding Systems 2*, pages 524–532. Morgan Kaufmann, Los Altos, CA, 1990.

[Fausett, 1994] Laurene Fausett. *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*. Prentice Hall, Englewood Cliffs, NJ, 1994.

[Fung and Enderwick, 2002] Eric T. Fung and Cynthia Enderwick. ProteinChip clinical proteomics: Computational challenges and solutions. *Computational Proteomics Supplement*, 32:S34–S41, 2002.

[Hlavaty *et al.*, 2001] John J. Hlavaty, Alan W. Partin, Felicity Kusinitz, Matthew J. Shue, Adam Stieg, Kate Bennett, and Joseph V. Briggman. Mass spectroscopy as a discovery tool for identifying serum markers for prostate cancer. *Clinical Chemistry*, 47(10):1924–1926, 2001.

[Issaq *et al.*, 2002] Haleem J. Issaq, Timothy D. Veenstra, Thomas P. Conrads, and Donna Felschow. The SELDI-TOF MS approach to proteomics: Protein profiling and biomarker identification. *Biochemical and Biophysical Research Communications*, 292(3):587–592, 2002.

[Li *et al.*, 2002] Jinong Li, Zhen Zhang, Jason Rosenzweig, Young Y. Wang, and Daniel W. Chan. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry*, 48(8):1296–1304, 2002.

[Paweletz *et al.*, 2000] Cloud P. Paweletz, John W. Gillespie, David K. Ornstein, Nicole L. Simone, Monica R. Brown, kristina A. Cole, Quan-Hong Wang, Jing Huang, Nan Hu, Tai-Tung Yip, William E. Rich, Elise C. Kohn, W. Marston Linehan, Thomas Weber, Phil Taylor, Mike R. Emmert-Buck, Lance A. Liotta, and Emanuel F. Petricoin. Rapid protein display profiling of cancer progression directly from human tissue using a protein biochip. *Drug Development Research*, 49:34–42, 2000.

[Petricoin and Liotta, 2002] Emanuel F. Petricoin and Lance A. Liotta. Proteomic analysis at the bedside: Early detection of cancer. *Trends in Biotechnology*, 20(12)Suppl.:S30–S34, 2002.

[Petricoin *et al.*, 2002a] Emanuel F. Petricoin, Ali M. Ardekani, Ben A. Hitt, Peter J. Levine, Vincent A. Fusaro, Seth M. Steinberg, Gordon B. Mills, Charles Simone, David A. Fishman, Elise C. Kohn, and Lance A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359:572–577, 2002.

[Petricoin *et al.*, 2002b] Emanuel F. Petricoin, David K. Ornstein, Cloud P. Paweletz, Ali Ardekani, Paul S. Hackett, Ben A. Hitt, Alfredo Velassco, Christian Trucco, Laura Wiegand, Kamillah Wood, Charles B. Simone, Peter J. Levine, W. Marston Linehan, Michael R. Emmert-Buck, Seth M. Steinberg, Elise C. Kohn, and Lance A. Liotta. Serum proteomic patterns for detection of prostate cancer. *Journal of the National Cancer Institute*, 94(20):1576–1578, 2002.

[Petricoin *et al.*, 2002c] Emanuel F. Petricoin, Kathryn C. Zoon, Elise C. Kohn, J. Carl Barrett, and Lance A. Liotta. Clinical proteomics: Translating benchside promise into bedside reality. *Nature Reviews Drug Discovery*, 1:683–695, 2002.

[Poon and Johnson, 2001] Terence C. W. Poon and Philip J. Johnson. Proteome analysis and its impact on the discovery of serological tumor markers. *Clinica Chimica Acta*, 313:231–239, 2001.

[Poon *et al.*, 2003] Terence C. W. Poon, Tai-Tung Yip, Anthony T. C. Chan, Christine Yip, Victor Yip, Tony S. Mok, Conrad C. Y. Lee, Thomas W. T. Leung, Stephen K. W. Ho, and Philip J. Johnson. Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. *Clinical Chemistry*, 49(5):752–760, 2003.

[Qu *et al.*, 2002] Yinsheng Qu, Bao-Ling Adam, Yutaka Yasui, Michael D. Ward, Lisa H. Cazares, Paul F. Schellhammer, Ziding Feng, Oliver John Semmes, and George L. Wright. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry*, 48(10):1835–1843, 2002.

[Quinlan, 1993] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[Sauter *et al.*, 2002] Edward R. Sauter, Weizhu Zhu, X.-J. Fan, R. P. Wassell, Inna Chervoneva, and Garrett C. Du Bois. Proteomic analysis of nipple aspirate fluid to detect biologic markers of breast cancer. *British Journal of Cancer*, 86:1440–1443, 2002.

[Srinivas *et al.*, 2001] Pothur R. Srinivas, Sudhir Srivastava, Sam Hanash, and George L. Wright. Proteomics in early detection of cancer. *Clinical Chemistry*, 47(10):1901–1911, 2001.

[Valerio *et al.*, 2001] A. Valerio, Daniela Basso, S. Mazza, G. Baldo, A. Tiengo, S. Pedrazzoli, R. Seraglia, and M. Plebani. Serum protein profiles of patients with pancreatic cancer and chronic pancreatitis: Searching for a diagnostic protein pattern. *Rapid Communications in Mass Spectrometry*, 15(24):2420–2425, 2001.

[Vlahou *et al.*, 2001] Antonia Vlahou, Paul F. Schellhammer, Savvas Mendrinos, Keyur Patel, Filippos I. Kondylis, Lei Gong, Suhail Nasim, and George L. Wright. Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. *American Journal of Pathology*, 158(4):1491–1502, 2001.

[Watkins *et al.*, 2001] Brynmor Watkins, Robert Szaro, Shannon Ball, Tatyana Knubovets, Joseph Briggman, John J. Hlavaty, Felicity Kusinitz, Adam Stieg, and Ying-Jye Wu. Detection of early-stage cancer by serum protein analysis. *American Laboratory*, 33:32–36, 2001.

[Wellmann *et al.*, 2002] Axel Wellmann, Volker Wollscheid, Hong Lu, Zhan Lu Ma, Peter Albers, Karin Schütze, Volker Rohde, Peter Behrens, Stefan Dreschers, Yon Ko, and Nicolas Wernert. Analysis of microdissected prostate tissue with ProteinChip arrays—a way to new insights into carcinogenesis and to diagnostic tools. *International Journal of Molecular Medicine*, 9:341–347, 2002.

[Wu *et al.*, 2003] Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 2003. To appear.

[Wulfkuhle *et al.*, 2001] Julia D. Wulfkuhle, Kelley C. McLean, Cloud P. Paweletz, Dennis C. Sgroi, Bruce J. Trock, Patricia S. Steeg, and Emanuel F. Petricoin. New approaches to proteomic analysis of breast cancer. *Proteomics*, 1(10):1205–1215, 2001.

[Wulfkuhle *et al.*, 2003] Julia D. Wulfkuhle, Lance A. Liotta, and Emanuel F. Petricoin. Proteomic applications for the early detection of cancer. *Nature Reviews Cancer*, 3:267–275, 2003.

[Yates, 2000] John R. Yates. Mass spectrometry: From genomics to proteomics. *Trends in Genetics*, 16(1):5–8, 2000.