APPLICATION OF THE RELEVANCE VECTOR MACHINE AND SUPPORT VECTOR MACHINE TO CLINICAL DATA

Elie Tcheimegni , Dr. Manohar Mareboyana,
Dr.Claude Turner
Department of Computer science , Bowie State University
14000 Jericho Park Road, Bowie, MD 20715
Elie Tcheimegni <tcheimegni@yahoo.fr>;
Claude F. Turner <cturner@bowiestate.edu>;
Manohar Mareboyana <mmareboyana@bowiestate.edu>;

Dr. Kofi Nyarko
Electrical & Computer Engineering (ECE)
Morgan State University, School of Engineering Authors
5200 Perring Parkway, Baltimore, MD 21251
Kofi Nyarko <kofi.nyarko@morgan.edu>;

Abstract—

The Relevance Vector Machine (RVM) algorithm has been widely utilized in many applications, such as machine learning, image pattern recognition, and compressed sensing .

Base on available training data from cancer and normal patients, the article presents and build a classifier suitable for genetic diagnosis, as well as drug discovery using sparse Bayesian learning , support vector machine (SVM), Relevance Vector Machine (RVM) to healthcare , and multilabel machine learning classification algorithms. Also, the computation procedure of the RVM algorithm is fully analyzed. Motivate by improvements of health diseases and cancers depiction that will be facilitated by an ability to predict the related syndrome occurrence , this work employs a data-driven approach to developing cancer prediction models using Relevance Vector Machine (RVM), a probabilistic kernel-based learning machine

The model is applied to cancer, tumor , and general health diseases . The results obtained using RVM are compared with those of state-of-the-art Support Vector Machine (SVM) to present the advantages of RVMs over SVMs. The finding results allow us to conclude that RVM is almost equal to SVM on training efficiency and classification accuracy, but RVM performs better on sparse property, generalization ability, and decision speed.

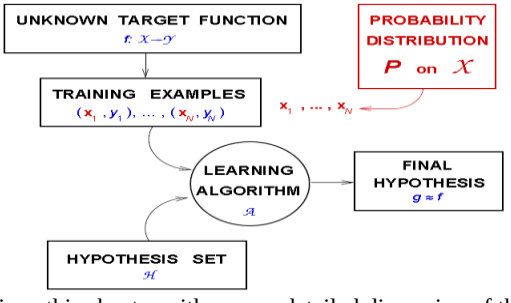
Keywords—component; formatting; style; styling; insert *(key words)* ***Support Vector Machines, Relevance Vector Machine, Rapidminer, Tanagra , Accuracy's values***

I. INTRODUCTION

Cancer is a group of diseases in which cells in the body grow, change, and multiply out of control(West, Mangiameli, Rampal, & West, 2005). Cancer can either be cancerous (malignant) or non-cancerous (benign). Malignant tumors penetrate and destroy healthy body tissues(Wolberg, Street, & Mangasarian, 1993) .Cancer detection has become a significant area of research in pattern recognition community. Each year millions of people including patients, caregivers, and health seekers use the World Wide Web (WWW, Internet) to retrieve health related information to meet their health care needs. Health care in the coming century requires the managing of public and private health concerns. Also, classifying health information for the purpose of data analysis , prediction and treatment of syndrome or seeking health information (either by using the Internet or other sources) is a strategy that many people use as a means of coping and reducing stress (Molen, 1999,(Akhu-zaheya, 2007)). Patients' insufficient health information is a factor that impedes them from participating in their treatment decisions (Gaston & Mitchell, 2005). It is, therefore, important to gather health information from patients, understand, or at least be able to classify and discriminate data obtained for the purpose of treatment of illness. However, for some type of diseases or tumors, this requires an understanding of the syndrome , which seems particularly difficult. Therefore, prediction methods have been developed as alternative ways to obtain that structure information.

The work of this thesis is the investigation and development of cancer and some common clinical diseases prediction capability in the medical field using the Relevance Vector Machine (RVM) and Support Vector Machine (SVM) .The methods and tools used for prediction are expected to have application to other regions of scientific research.

Fig. 1 shows the various stages followed for the design of a classification system. As it is apparent from the feedback arrows, these stages are dependent. On the contrary, they are interrelated and, depending on the results, learning algorithm and hypothesis set, one may go back to redesign earlier stages in order to improve the overall performance by modifying the hypothesis set or learning algorithm.



We continue this chapter with a more detailed discussion of the situation in medical field, explaining how the available data lends itself to a learning machine approach.

A. SUPPORT VECTOR MACHINE (SVM): AN OVERVIEW

In the supervised learning paradigm, training data is comprised of pairs of input/output points $\{(x_i, y_i)\}_{i=1}^n$, where y_i is a finite discrete or continue variable indicating the label associate to x_i, with x_i belonging to some space $X = \Omega$ and $y_i \in IR$ (regression) or $y_i \in \{-1,1\}$ (binary classification). A classification model is constructed from a set of data for which the attributes and the true classes are known and is employed to assign a class to a new object on the basis of its observed attributes or features.

An advantage of this algorithm is its sparsity since only a small subset of the training samples are finally retained for the classifier.

B. Relevance Vector Machine oVERVIEW

As a supervised learning, RVM starts with a set of data inputs $\{x_i\}_{i=1}^n$ and their corresponding target vectors $\{t_i\}_{i=1}^n$. The aim is to learn a model of the dependency of the target vectors on the inputs in order to make accurate prediction of t for unseen value of x. Typically, the predictions are based on a function *y*(x) defined over the input space, and learning the process of inferring the parameter of this function. The value $t_* = y(x_*)$ of a function $y(x)$ needs to be predicted at some arbitrary point x_*, given a set of (typically noisy) measurements of the function $\{t_i\}_{i=1}^n$ at some training points $\{x_i\}_{i=1}^n$, $t_i = y(x_i) + \varepsilon_i$ where ε_i is the noise component of the measurement (Tzikas, Wei, Likas, Yang, & Galatsanos, n.d.). Under a linear model assumption, the unknown function $y(x)$ is a linear combination of some known basis functions $\phi_i(x)$, i.e. $y(x) = \sum_{i=1}^m w_i \phi_i(x)$, where $w = (w_1, w_2, ..., w_m)$ is a vector consisting of linear combination of weights , so $t = \Phi w + \varepsilon$, where Φ is an $n \times m$ matrix whose it-yh colum is formed with the values of basis function $\phi_i(x)$ at all the training points, and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ is the noise vector;we will assume independent, zero-mean, Gaussian distribution for the noise term, i.e. $\varepsilon_i \sim N(0, \sigma^2)$. In the context of SVM, this function takes form $y(x) = \sum_{i=1}^n w_i k(x, x_i) + w_0$, w_0 is the bias, where $w = (w_1, w_2, ..., w_n)$ is weight vectors, w_0 is bias and $k(x, x_i)$ is a kernel function, is the basis function vector.

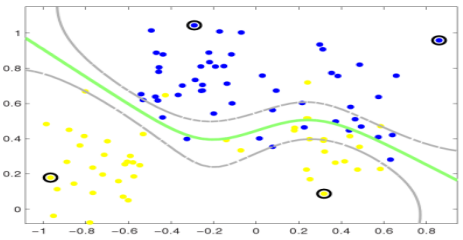
Given a feature vector x, our goal is to perform classification based on the probability that v is associated with hypotheses H_0 (false target) and H_1 (true target), denoted respectively by $p(H_0|x)$ and $p(H_1|x)$ (Fletcher, 2010). Assume access to ‘‘training’’ data $\{x_i, t_i\}_{i=1}^n$, where x_i represent feature vectors and t_i represent scalar labels. In the regression problem the objective is to learn a function that predicts the label of a given feature vector. Predictor function *f* can be written as $f(x, w) = \sum_{i=1}^n w_i \phi(x, x_i) + w_0$ where $\phi(x, x_i)$ is a kernel that quantifies the similarity between feature vectors x and x_i . Any linear or nonlinear measure may be selected for the kernel. It is assumed that the error ε_i between the regression function $f(x_i, w)$ and true label t_i is an independent, identically distributed (iid) zero-mean Gaussian with variance σ^2 , yielding $\varepsilon_i = t_i - f(x_i, w) \Leftrightarrow t_i = f(x_i, w) + \varepsilon_i$, and for the n training examples the likelihood of the label set L can be represented as

$$p(L|w, \sigma^2) = p(T|w, \sigma^2) = (2\pi\sigma^2)^{\frac{-n}{2}} \exp\left(-\frac{\|T - \Phi w\|^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{\frac{-n}{2}} \exp\left(-\frac{\|L - \Phi w\|^2}{2\sigma^2}\right)$$

where $L = T = \{t_1, t_2, ..., t_n\}$, $W = \{w_0, w_1, ..., w_n\}$, and Φ is a kernel design matrix of size $n \times (n+1)$ whose ith row $\Phi_i = [1, \phi(x_i, x_1), ..., \phi(x_i, x_n)]$. The likelihood of data set can be written as $p(T|w, \sigma^2) = (2\pi\sigma^2)^{n-2} \exp\left(-\frac{1}{2\sigma^2}\|T - \Phi w\|^2\right)$ where Φ is the $n \times (n+1)$ design matrix with $\Phi_{n(n+1)} = \{1, K(x_t, x_1), K(x_t, x_2), , K(x_t, x_n)\}^T$.

To control the complexity of model and avoid overfitting, a zero-mean Gaussian prior probability distribution is defined(Tipping, 2001) over every $\underline{\omega_i}$ with variance σ_i^{-1} , the likelihood of *W* is written as: $p(w|\alpha) = (2\pi)^{\frac{-n}{2}} \prod_{i=1}^n \alpha_i^{\frac{1}{2}} \exp\left(-\frac{\alpha_i w_i^2}{2}\right)$ where hyperparameters vector $\alpha = (\alpha_0, \alpha_1, ..., \alpha_n)^T$, controls how far from zero each weight is allowed to deviate. To control the model sparseness, RVM defined a hierarchical prior over α : $p(\alpha)$ and the inverse noise variance σ^2: $p(\sigma^2)$ is specified uninformative hyperpriors: Gamma distributions. With the addition of noise $\varepsilon_i : t_i = y_i + \varepsilon_i = w^T \phi(x_i) + \varepsilon_i$ where ε_i are assumed to be independent samples from a Gaussian noise process with zero mean and variance σ^2 ,i.e. $\varepsilon_i \sim N(0, \sigma^2), \forall i$.



RVM two-class classification example with four relevance vectors(Silva & Ribeiro, 2010).

As Tipping (2000), to avoid overfitting problems which may be caused by the Maximum likelihood estimation of *w* and σ^2 , zero mean Gaussian prior over the weights *w* is introduced. The prior on the weights is independent Gaussian, $P(w|\alpha_i) \sim N(0, \alpha_i^{-1})$ where we have used α_i to describe the inverse variance (i.e. precision) of each w_i .

To make the RVM favor sparse regression models, Tipping add a new parameter $\alpha_i \in (0;+\infty)$ for each weight parameter w_i . The parameter α_i affects to w_i through a conditional distribution $P(w_i|\alpha_i)$, which is given by $p(w_i|\alpha_i) = N(0, \alpha_i^{-1})$ where the notation $N(B|\mu, \sigma^2)$ specifies a Gaussian distribution over B with mean μ and variance σ^2 . Variables such as μ and σ^2 are sometimes called *hyperparameters* since they control the distribution over parameters.

The likelihood of the weights w, hyperparameters α , and variance σ^2 given the training data is expressed as $p(w, \alpha, \sigma^2|T) = p(w|\alpha, \sigma^2, T)p(\alpha, \sigma^2|T)$. The first term on the right can be written as $p(w|\alpha, \sigma^2, T) = \frac{p(T|w, \sigma^2)p(w|\alpha)}{p(T|\alpha, \sigma^2)}$ By choosing $p(w|\alpha)$ as a product of Gaussians, $p(w|\alpha) = \prod_{i=0}^{n} N(w_i|0, \alpha_i^{-1})$ (this is a Gaussian prior), and $P(T|w, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\|T - \Phi W\|^2}{2\sigma^2}\right\}$ is also Gaussian likelihood implying the posterior $p(w|t, \alpha, \sigma^2)$ is Gaussian which obtained as (Tipping, 2004), the integral $p(T|\alpha, \sigma^2) = \int p(T|w, \sigma^2)p(w|\alpha)dw$ can be evaluated analytically (it is a convolution of Gaussians).

The posterior distribution over the weights is thus given by : $p(w|\alpha, \sigma^2, T) = \frac{p(T|w, \sigma^2)p(w|\alpha)}{p(T|\alpha, \sigma^2)}$

$$= (2\pi)^{\frac{(n+1)}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{(w-\mu)^T \Sigma^{-1}(w-\mu)}{2}\right\}$$

$$p(w|t, \alpha, \sigma^2) = (2\pi)^{\frac{(n+1)}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{(w-\mu)^T \Sigma^{-1}(w-\mu)}{2}\right\}$$

with covariance $\Sigma = (\sigma^{-2}\Phi^T\Phi + A)^{-1}$ and mean $\mu = \sigma^{-1}\Sigma\Phi^T t$ where $A = diag(\alpha_0, \alpha_1, ..., \alpha_n)$

The equation $p(w|t, \alpha, \sigma^2) = \frac{p(t|w, \sigma^2)p(w, \alpha)}{p(t|\alpha, \sigma^2)}$ has an analytical solution where the posterior covariance and mean are $\Sigma = (\Phi^T B\Phi + A)^{-1}$, $\mu = \Sigma\Phi^T Bt$ with A = *diag*(α_1, … , α_{N+1}), and $B = \sigma^{-2}I$.

The evaluation of $p(\alpha, \sigma^2|L) \propto p(L|\alpha, \sigma^2)p(\alpha)p(\sigma^2)$ requires introduction of prior distributions $p(\alpha)$ and $p(\sigma^2)$. Since α and σ^2 are associated with the variances of Gaussian distributions, they must be purely positive. Therefore gamma priors are introduced

$$p(\alpha) = \prod_{i=0}^{N} \text{Gamma}(\alpha_i|\text{a}, \text{b}); \; p(\beta) = \text{Gamma}(\beta|\text{c}, \text{d}) \quad (11)$$

wit $\beta = \sigma^{-2}$ and $\text{Gamma}(\alpha_i|\text{a}, \text{b}) = (\Gamma(a))^{-1} b^a \alpha_i^{a-1} e^{-b\alpha_i}$, **with** $\Gamma(a) = \int_0^{+\infty} x^{a-1}e^{-x}dx$. The parameters *a*, *b*, *c*, and *d* are set to zero, yielding a set of uniform hyperpriors(Jeffrey's prior). The expression $p(\alpha, \sigma^2|L)$ cannot be evaluated analytically, and therefore it is approximated as a delta function at its mode: $p(\alpha, \sigma^2|L) \approx \delta(\alpha - \alpha_{MP})\delta(\sigma^2 - \sigma^2_{MP})$. The most probable α_{MP} and σ^2_{MP} are computed by maximizing $p(\alpha, \sigma^2|L)$, **which** for the uniform hyperparameters reduces to maximizing $p(L|\alpha, \sigma^2) = \int p(L|w, \sigma^2)p(w|\alpha)dw =$ The

$$(2\pi)^{-\frac{N}{2}} |\sigma^2 I + \Phi A^{-1}\Phi^T|^{-\frac{1}{2}}$$

$$\times \exp\left[-\frac{1}{2}L^T(\sigma^2 I + \Phi A^{-1}\Phi^T)^{-1}L\right]$$

maximization yields $\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2}$ where

$\gamma_i \equiv 1-\alpha_i\Sigma_{ii}$ And $(\sigma^2)^{new} = \frac{\|L-\Phi\mu\|^2}{N-\Sigma_i\gamma_i}$ The learning process is therefore characterized by repeated application of $\gamma_i \equiv 1-\alpha_i\Sigma_{ii}$ and $(\sigma^2)^{new} = \frac{\|L-\Phi\mu\|^2}{N-\Sigma_i\gamma_i}$,concurrent with updating $\Sigma = (\sigma^{-2}\Phi^T\Phi + A)^{-1}$ and $\mu = \sigma^{-2}\Sigma\Phi^T L$ .The composite prior distribution on the weights, defined

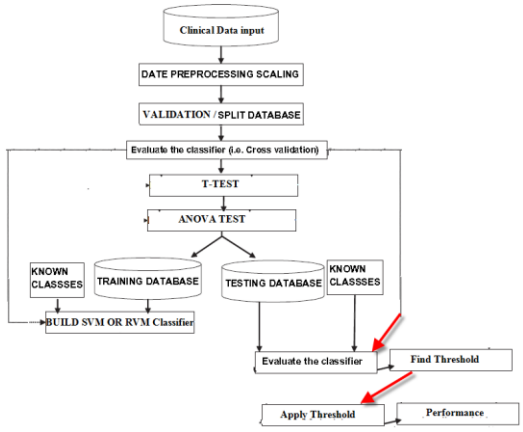
$$p(w|\alpha) = \prod_{i=0}^{N} N(w_i|0,\alpha_i^{-1})$$

By and $p(\alpha) = \prod_{i=0}^{N} Gamma(\alpha_i|a,b)$ yields an algorithm in which most of the α_i in $\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2}$, $\gamma_i \equiv 1-\alpha_i\Sigma_{ii}$ go to ''infinity'' (reaching the maximum precision of the computer), and the associated weight w_i has a likelihood highly peaked at zero. The associated feature vector x_i is deemed ''irrelevant'' to the function $f(x,w)$ that one wishes to learn, and the irrelevant feature vectors are pruned via iterative update of $\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2}$, $\gamma_i \equiv 1-\alpha_i\Sigma_{ii}$ and $(\sigma^2)^{new} = \frac{\|L-\Phi\mu\|^2}{N-\Sigma_i\gamma_i}$. The algorithm therefore works by seeking to maximize the likelihood $p(\alpha,\sigma^2|L)$, under the constraint imposed by the prior distributions $p(w|\alpha)$ and $p(\alpha)$. Once the most-probable parameters α_{MP} and σ^2_{MP} are estimated within the training phase, the joint posterior of model parameters is approximated as

$$p(w,\alpha,\sigma^2|L) = p(w|\alpha,\sigma^2,L)p(\alpha,\sigma^2|L) \approx p(w|\alpha_{MP},\sigma^2_{MP},L)$$

II. METHODOLOGY

In order to classify different clinical disease data, . It is necessary to understand the medical nature of the disease in order to construct or to find the best learning system for the particular disease. Information of the database describes where the database was obtained, the size of the database, and whether the database contains samples with missing value. All the medical databases are obtained from the machine learning repository, Department of Information and Computer Science, University of California at Irvine(Asuncion & Newman, 2007). We will use the following flowchart to classify data



-The breast cancer database used here where obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg (Mangasarian, O. L. and Wolberg, 1990). The database contains 699 samples with 683 complete data and 16 samples with missing attributes. There are 9 integer-valued attributes and each data values range from 1 to 10. These attributes measure the external appearance and internal chromosome changes in nine different scales. There are two values in the class variable of breast cancer: benign (non-cancerous) and malignant (cancerous), which is represented numerically by 0 and 1 respectively.

-The liver disorders database was donated by Richard S. Forsyth, BUPA Medical Research Ltd. There are 345 samples in total and they are all complete data. It contains 6 attributes and the first five attributes are blood tests. All blood tests are to be sensitive to the liver disorders that might arise from excessive alcohol consumption.

-The Pima Indian Diabetes Database was obtained from US National Institute of Diabetes and Digestive and Kidney Diseases. All patients are female, with Pima Indian heritage, who live near Phoenix, Arizona, USA. There are 768 data samples and all samples have no missing attributes. There are 8 attributes in this database. There are only two output classes, diabetes and non-diabetes. It is a typically difficult medical diagnosis problem where only a small amount of database is available. Splitting data is a very important process to be done before begin to predict the data. The reason to split data is because there is a need to identify the best split technique in order to get the higher accuracy for the model. The methodology of measuring performance is divided into two major types: single training/testing data scheme and random sub-sampling scheme. Single training/testing data scheme is not a reliable estimator of the classification performance on a small database especially when the size of the database is less than several thousand. Due to all the medical databases are

relatively small, and hence single training/testing data scheme is not the mainstream classification methodology in this thesis.

The second type of methodology is random sub-sampling scheme, and the main idea is by selecting a random training/testing data in the database to minimize the classification bias. In general, there are two main kinds of random sub-sampling scheme: a leave-one-out cross-validation and N-fold cross-validation; and both kinds of scheme are described by Kohavi in 1995(Cheung, 2001; Kohavi, 1995). In 1983, Efron claimed that leave-one-out cross-validation gives nearly unbiased estimates of the accuracy, however it often returns with unacceptably high variables especially for small databases(Kohavi, 1995) . In the experiment, we will import a dataset and train a support vector machine model or Relevance vector model. We will use cross validation to evaluate the accuracy of our learning model. Cross validation works by using part of the data to train the model, and the rest of the dataset to test the accuracy of the trained model. As previously described, all the medical databases are relatively small, therefore, in this thesis, the classification methodology concentrateson the N-fold cross-validation scheme and considers what number of N to use in this scheme. In general, N is chosen either 5 or 10 due to computational complexity which may increase significantly if N is larger than 10. In 1994, Weiss and Indurhkya indicated that for database's size of at least 200, using N =10 to choose the amount of pruning unbiased trees is optimal(Kohavi, 1995)

EXPERIMENT

Table below shows the accuracy of different classification methods, and each classification method is described throughout the thesis. It also highlights the best classification method in bold.

SVM-RVM prediction Accuracy

CSVC,RVM with RBF kernel with $c = 1.5, \gamma = 0.5$ convergence epsilon 0.0010; kernel gamma; kernel lengthscale :3.0

Database	SVM accuracy	SVM sensitivity	SVM error	RVM accuracy	RVM sensitivity	RVM error
Breast Cancer Wisconsin	65.52% +/- 0.46%	0.83% +/- 1.67%	34.48% +/- 0.46%	65.52% +/- 0.46%	0.00% +/- 0.00%	34.48% +/- 0.46%
Liver Disorders	64.25% +/- 3.68%	38.25% +/- 9.26%	35.75% +/- 3.68%	57.54% +/- 8.04%	66.75% +/- 16.87%	42.46% +/- 8.04%
Pima Indians Diabetes	34.24% +/- 1.21%	3.80% +/- 1.94%	65.76% +/- 1.21%	34.90% +/- 0.21%	0.00% +/- 0.00%	65.10% +/- 0.21%

False positive

CSVC,RBF kernel with $c = 1.5, \gamma = 0.5$ convergence epsilon 0.0010; **kernel gamma** ; **kernel lengthscale** =3.0

Database	**SVM** false negative	**SVM false positive**	**SVM AUC**	**RVM** false negative	**RVM false positive**	**RVM AUC**
Breast Cancer Wisconsin	47.800 +/- 1.166	0.400 +/- 0.800	0.502 +/- 0.004	48.200 +/- 0.748	0.000 +/- 0.000	0.000 +/- 0.000
Liver Disorders	17.800 +/- 2.786	6.800 +/- 0.748	0.641 +/- 0.053	9.600 +/- 4.923	19.600 +/- 10.288	0.597 +/- 0.025
Pima Indians Diabetes	96.200 +/- 1.939	4.800 +/- 1.600	0.423 +/- 0.029	100.000 +/- 0.000	0.000 +/- 0.000	0.508 +/- 0.029

Results and Discussion

We have proven that it's possible to develop a kernel method without using kernel trick which can equal or perform better than SVM. One of the most appealing features of kernel algorithms is the solid foundation provided by both statistical learning theory and functional analysis. Kernel methods let us interpret (and design) learning algorithms geometrically in feature spaces non-linearly related to the input space, and combine statistics and geometry in a promising way. Support vector machines have been one of the major kernel methods for data classification. Its original form requires a parameter $C \in [0;+\infty)$; which controls the trade-off between the classifier capacity and the training errors.

The methods presented in this thesis were developed based on a new statistical learning machine tool, the Relevance Vector Machines. We apply the relevance vector machine algorithm experimentally to several benchmark sets for classification problem. The advantages of this model is its probabilistic approach to provide a mapping function among the available input-output data, while avoiding overfitting issues that could affect its performance, which have been often seen in previous engineering applications of data-driven models. The Bayesian framework of the RVM model features sparsity, accuracy, and incorporation statistical judgment of the posterior probability; it uses this to produce probabilistic output. In general, the results for each of the methods developed were satisfactory. By contrasting to SVM, we draw this conclusions: RVM has comparable classification accuracy to SVM, but it's much sparser; the number of relevance vector increases slower than that of support vector when the size of training set grows; applied to unlearned samples, the generalization ability of RVM is better than that of SVM overall; finally, as RVM is sparser, it's faster than SVM on decision speed. RVM algorithms have the advantage of being more robust and stable than the alternative, the SVM, as was found in each analysis

.SVM can handle nonlinear relationship efficiently by implicitly transforming the input space into another higher dimensional space. However, SVM is not favorable for huge datasets training**.**

Future Research

Not least important, the ability to apply RVM technique to real-life problems is essential to its survival. how to actively embed prior knowledge of the problem at hand within RVM, in a probabilistic Bayesian approach can be considered an immediate goal. The relationship between the process of feature selection and whether(or when) it make sense to perform the latter is an open question that should be investigate. In future some

advanced neural network techniques can be used to train the RVM classifier and it may enhance the classification accuracy of medical data and reduce the training time.

In this thesis we have offered several applications of the technique to real-life problems where we have matched or exceeded the state -of-the art benchmarks. We believe there will be more to follow.

Conclusion

This thesis marks the conclusion of four years of hard work and research surround an excellent, and very promising tool. Relevance vector machine is a probabilistic kernel-based learning machine method. The results confirm that the RVM classification system substantially boosts the generalization capability achievable with the SVM classifier. Another advantage of the RVM approach can be found in its high sparseness, which is explained by the fact that the adopted optimization criterion is based on minimizing the number of support vectors. We believe that Relevance Vector Machine will give many researchers an outstanding benchmark with data mining in the area of Bioinformatics beside Support Vector Machine and Neuron Network. . It's elegance, probabilistic approach, simplicity and sound mathematical framework make RVMs a very unique tool within the reach of all machine learning researchers. Undoubtly, there will be more to follow.

References

Akhu-zaheya, L. M. (2007). Factors Influencing Health Information-Seeking Behavior of Jordanian Patients With Cancer by Faculty of the Graduate School of the State University of New York at Buffalo in partial fulfillment of the requirements for the.

Asuncion, A., & Newman, D. J. (2007). UCI Machine Learning Repository. *University of California Irvine School of Information*. University of California, Irvine, School of Information and Computer Sciences. Retrieved from http://www.ics.uci.edu/~mlearn/MLRepository.html

Cheung, N. (2001). Machine Learning Techniques for Medical Analysis, (October).

Fletcher, T. (2010). Relevance Vector Machines Explained.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*, *14*(12), 1137–1143. doi:10.1067/mod.2000.109031

Mangasarian, O. L. and Wolberg, W. H. (1990). Cancer Diagnosis via linear programming. *SIAM News, vol. 23*(1990), pp. 1–18.

Silva, C., & Ribeiro, B. (2010). Inductive Inference for Large Scale Text Classification. *Silva*, *255*, 117–128. doi:10.1007/978-3-642-04533-2

Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *J ournal of Machine Learning Research*.

Tzikas, D. G., Wei, L., Likas, A., Yang, Y., & Galatsanos, P. (n.d.). A tutorial on relevance vector machines for regression and classification with applications.

West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal Of Operational Research*, *162*(2), 532–551. doi:10.1016/j.ejor.2003.10.013

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1993). Breast cytology diagnosis with digital image analysis. *Analytical and quantitative cytology and histology the International Academy of Cytology and*

American Society of Cytology, *15*(6), 396–404. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8297430