2009 Ninth International Conference on Intelligent Systems Design and Applications

Profit Forecasting Using Support Vector Regression for Consulting Engineering Firms

Víctor Yepes, Eugenio Pellicer
School of Civil Engineering
Universitat Politècnica de València
C. de Vera s/n, 46022 Valencia, Spain
{vyepes,pellicer}@cst.upv.es

Francesc J. Ferri
Computer Science Department
Universitat de València
Dr Moliner 50, 46100 Burjassot, Spain
francesc.ferri@uv.es

***Abstract*—This paper introduces Support Vector Machines (SVM) in the particular field of decision support systems for consulting engineering companies and studies the differences and particularities of the corresponding solutions. A detailed analysis has been performed in order to assess the suitability and adaptability of these methods for the particular task taking into account the risk/benefit tradeoff.**

***Index Terms*—Support Vector Machines; Pattern Classification; Decision Support Systems; Consulting; Bidding;**

I. Introduction

Construction can be characterized, in fact, as a process with five main phases: feasibility, design, construction, operation, and divestment. For each phase, different types of projects are required and, as a result, specialized project-based companies perform, being consulting engineering firms and contractors the most widespread. Generally, consulting engineering companies take care of the first two phases of the process (feasibility and design); these companies yield very specialized technical knowledge into outputs: appraisal reports, design projects or inspections [1]. Thus, consulting engineering firms work to contracts awarded by the clients. Cost and price are the two variables that operate in order to obtain the final economic result (profit). Many times, price is fixed before the job starts; therefore profit depends on the cost of the project. This is especially true in some fields like civil engineering, where clients are governmental agencies mostly. These clients tend to procure contracts based mainly on the price submitted in sealed bids by companies (closed bidding).

For that reason, the bidding process is crucial for consulting engineering firms working in the construction industry, remarkably in the sub-sector of civil engineering. This is also applicable to contractors. In both cases, the turnover of the company depends very much on the success of these closed biddings. Thus, two different but sequential stages of bidding decision can be distinguished: bid/no bid; and mark-up [2]. Determining the economic bid made to the client is one of the most important decisions that these kinds of companies have to make [3]. If the price offered is high, then the probability of the contract being awarded to the company is very low, mostly when the competition is significant. In contrast, if the price submitted is low, it is easier to win the contract. Nonetheless, if the price is too low and the company is awarded the contract, it is likely that the resulting profit is low. To the point that it could even generate important losses to the company.

Besides, the macroeconomic context also shapes in these decisions. When the demand is important and the offer is low (recessive scenarios), competition is so hard that it can force to bid very low prices; in this case, profit is scarce or it does not appear (economic losses). In case the company is awarded the contract, the decision makers problem is to choose what bids mark-up would result in economic losses. Stated differently, if the price offered is fixed because of the market context, what is the probability of having profit or losses?. In addition, each kind of client or project could influence differently the profitability of the contract.

This problem has already been addressed by some authors in the past, mainly from the point of view of contractors. The seminal work of Friedman [4] started the interest in the area of construction management, coining the expression competitive bidding strategy; he also proposed a simple decision support model. Gates [5] established a pricing model for tendering that became later an economic theory for pricing construction projects. Other authors analyzed many previous papers and criticized the lack of a suitable theoretical framework for bidding strategy. Researchers have applied different techniques in order to enlighten this problem. Ahmad [6] employed traditional techniques of decision analysis. Drew and Skitmore [3] used multiple regressions to construct a prediction model, exploiting bidder, contract type and contract size as independent variables. Fayek [7] brought into play fuzzy set theory to obtain qualitative data. Li and Love [8] integrated a rule-based expert system and an artificial neural network to support decisions in bidding estimation. Wanous et al. [9] drew on a simple parametric solution that took into account 38 factors. Chua et al. [10] presented a case-based reasoning system. Egemen and Mohamed [2] also used a reasoning model, identifying key determining factors.

This paper follows the path started by Li and Love [8] using an artificial neural network that seeks to support decision making in the company regarding strategic bidding. Data considered in this work comes from a Spanish consulting civil

This work has been partially funded by Feder and Spanish MEC projects DPI2006-15542-C04-04, TIN2009-14205-C04-03 and Consolider Ingenio 2010 CSD2007-00018.

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.43

engineering company and consequently it is restricted to a particular local context. This kind of companies can be easily characterized by two features: turnover and personnel [11]; for this particular firm, these values are 7.5 million of Euros and 110 employees, respectively, in 2007. Moreover, the firm is a traditional civil engineering focused company: 80% of the contracts are civil engineering projects for public agencies, whereas the rest comprises urban planning, industrial engineering, architecture or Research and Development.

II. Profit prediction as a machine learning problem

As in most previous works in the literature, the final goal would be to setup an appropriate decision support system that is able to guide the bidding process so that the profit for the company is maximized. Nevertheless, in this work a very related subgoal consisting on profit prediction once the project has been assigned is considered.

The motivation is twofold. On one hand, the problem can be tackled in this way within the inductive learning paradigm. On the other hand, the solution will be adapted for a particular firm in a particular context. Consequently, objective measures of performance will be derived which will serve to assess the proposed system.

The data which is available for this study has a very heterogeneous nature. Both categorical and numerical data has been collected. Moreover, the profit to be predicted (which is the real profit achieved for the considered projects) has an unpredictable, random component. In fact, for a consulting company it is not so important to exactly predict the profit to be obtained as long as you can roughly predict whether it is going to be positive or negative.

In particular, the input data obtained from the pilot company comprises 263 series of data. Each of the series stands for one contract of the firm, from 2001 to 2007. 31 input variables are considered, grouped as follows:

Type of contract (4): type of work, specialization (generic and specific), and client

Bidding data (4): budget, number of bidders, average bid price, and price awarded

Company forecast (3): duration (in months), time (in hours), and subcontracting (in Euros)

Actual data (6): starting year, actual duration, invoicing, total cost, technicians (in hours), and subcontracting (in Euros)

Macroeconomic data (13): gross domestic product (GDP) per capita, increase of GDP, gross fixed capital formation in construction, unemployment rate, consumer price index, gross debt of public agencies, synthetic interest bank rate, Madrid stock market index, exchange rate euro dollar, bidding of contracts (public works and services), average awarding price of contracts (public works and services), bidders per contract (public works and services), and production in construction. These data were obtained from INE (2008), BCC (2008), SEOPAN (2008), TECNIBERIA (2008), and BANCO DE ESPAÑA (2008).

Profit (1): the resulting profit for the company. This is the only output variable considered in this work.

III. Support vector regression

Support vector machines (SVM) are well-founded and widely employed classification and regression methods. From a technical point of view, SVM are linear predictors that operate in an appropriately transformed space that allows a wide range of nonlinear function possibilities as in the case of Multilayer Perceptrons. This nonlinear transformation of the input data is achieved by the so-called kernel trick [12].

The training of SVM is carried out by the structural risk minimization principle that in this particular case leads to a margin maximization problem that is solved by quadratic programming. In the particular case of regression, the process can be seen as finding the flattest linear function that approximates all training samples within a given approximation level. In other words, find the function $wx+b$ that

$$\begin{array}{l}\text{minimizes } \frac{1}{2}||w||^2\\ \text{subject to } |y_i - wx_i - b| \leq \varepsilon\end{array}$$

where $\{(x_i, y_i)\}_{i=1}^{\ell}$ are the available training examples and ε is the approximation level. The zone of width ε around the linear function is usually referred to as the insensitivity tube.

A conveniently softened version of the above problem is obtained by introducing penalty terms for samples outside the tube and slack variables that control these both at the positive and negative sides of the function. In particular the standard formulation is

$$\begin{array}{l}\text{minimize } \frac{1}{2}||w||^2 + C\sum(\xi_i + \xi_i^*)\\ \text{subject to } \left\{\begin{array}{l} y_i - wx_i - b \leq \varepsilon + \xi_i\\ wx_i + b - y_i \leq \varepsilon + \xi_i^*\\ \xi_i, \xi_i^* \geq 0\end{array}\right.\end{array} \tag{1}$$

In this formulation, the prespecified constant C determines the tradeoff between smoothnes and sensitivity to outliers while the value of ε specifies the range of deviations that will not be penalized. Solving the problem would mean obtaining the parameters of the linear function, w and b and corresponding values for each slack variable. This problem is solved in its dual form in terms of the Lagrange multipliers, $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$, for each of constraints in Eq. (1). In particular, both w and b can be written in terms of the variables α_i and α_i^* that can be shown to vanish for all examples inside the ε-insensitivity tube. The remaining examples x_i whose corresponding variables do not vanish are referred to as support vectors (SV) and the corresponding linear approximating function is the so-called Support Vector expansion [12], [13].

$$f(x) = \sum_{i=1}^{\ell}(\alpha_i - {\alpha_i}^*)x_i \cdot x + b$$

Figure 1 shows an illustrative example of the kind of linear approximation that is obtained for appropriate values of ε and C.

As all this formulation can be expressed exclusively in terms of dot products between elements (support vectors) in the original representation space, the method can be extended to the nonlinear approximation case using kernel functions. In particular, let ϕ be a nonlinear mapping from the original space to a high, possibly infinite dimensional space in which dot products can be computed as

$$\phi(x) \cdot \phi(x') = K(x, x')$$

where K is the so-called kernel function. Then the above linear approximation can be implicitly done in the target space which is nonlinearly related to the original one.

From all possible families of valid kernels, the radial basis function (RBF) or proximity kernel parameterized by an influence parameter, γ, is one of the most widely used becuse its close relation to nonparametric distance-based learning methods.

$$K(x, x') = e^{\gamma \cdot ||x-x'||^2}$$

For specific applications, particular distance functions can be used instead of the Euclidean norm as shown above [13], [14].

IV. Assessing learning results

Even though the problem of profit prediction has been posed as a regression one, it will be also interesting to assess the resulting predictors as two-class classifiers (profit/non profit) by applying a step-function on the SVR output. By considering a convenient range of threshold values, a family of different classifiers is obtained each of them with different error rates on both positive (profit) and negative (non profit) examples. The Receiver Operating Characteristic (ROC) curve is defined as the plot of the true positive rate (TP) against false positive rate (FP) considering the threshold used in the classifier as a parameter. The so-called ROC space is given by all possible results of such a classifier in the form of pairs (FP,TP). The performance of any classifier (with the corresponding threshold) can be represented by a point in the ROC space.

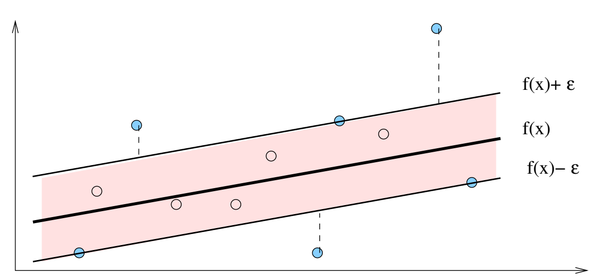

Fig. 1. Robust linear approximation obtained through the SV expansion. Support vectors are shown as solid dots. The shaded region represents the ε-insensitivity tube.

ROC curves move from the all-inactive point (0,0) which corresponds to the highest value of the threshold to the all-active point (1,1) given by the lowest value for the threshold. The straight line between these two trivial points in the ROC space corresponds to the family of random classifiers with different a priori probabilities for each class. The more a ROC curve separates from this line, the better the corresponding classification scheme is. As ROC curves move away from this line, they approach the best possible particular result that corresponds to the point (0,1) in the ROC space which means no false alarms and highest possible accuracy in the active class.

The ROC curve is a perfect tool to find the best trade-off between true positives and false positives and to compare classifiers in a range of different situations. A number techniques to obtain different measures from ROC curves have been also developed [15]. In particular, the Area Under the Curve (AUC) is a widely used measure that globally characterizes the performance of a given model througout the whole range of its threshold parameter.

V. Data preparation, normalization and experiments

The experiments presented in this work have been specifically designed in order to assess the applicability of SVR in the particular context of profit prediction given the available data from a particular consulting firm. As explained in Section II, only historical data corresponding to project details already completed by the firm are taken into account.

Nominal attributes have been converted into an appropriate number of numerical values taken into account that they will be equidistant of each other. Then all resulting attributes have been linearly normalized in the range $[0, 1]$ to avoid dominance problems when using a distance based kernel.

One of the main well-known problems of SV-based learning methods is parameter tuning. In the case of SVR with Gaussian kernels three parameters have to be taken into account. On one hand, the influence parameter of the kernel, γ that controls how distances between examples are taken into account in the representation space. On the other hand, the parameter C is related to the tradeoff between flatness or smoothnes of the solution and how deviations of particular points are to be penalized. Finally, the sensitivity parameter, ε that in our case could be easily translated into money, decides what amount of deviation is to be penalized or not.

Given the amount of data available and also because it has been decided (for this particular work) not to use any prior knowledge, the prediction problem is considerably hard. Within this context, both parameter tuning and performance assessment has been done using classification measures. In particular, both positive (profit) and negative (non profit) accuracy rates have been taken into account along with their geometric mean [16] in order to evaluate the the merits of the proposed approach.

In order to obtain results as significant as possible, 10-fold stratified cross-validation has been used to compute all

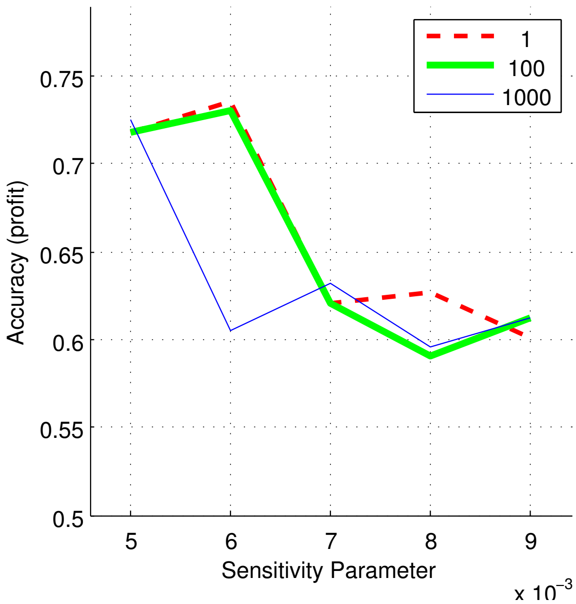

Fig. 2. Positive (profit) accuracy rate obtained for some of the experiments carried out. The influence parameter of the RBF kernel was fixed at $\gamma = 0.5$.

mesures of performance.

A wide range of parameter values has been tried when learning SVR models using random partitions of the available data. Surprisingly enough, the resulting models were relatively insensitive to these parameters from the point of view of predicting profit/non profit.

Figure 2 shows the results of part of the experiments performed for parameter setting. For these experiments, only the positive error rate is shown. Negative error rate remained almost constant slightly above 0.5 for the range of parameters shown. The value of the kernel parameter that was selected was $\gamma = 0.05$. The setting $C = 100$ and $\varepsilon = 0.006$ were also adopted based on the above experimentation.

Once parameters have been fixed, several experiments varying the number of training examples have been also performed in order to study how the method improves as more information is available.

Figure 3 shows a particular example of a SVR model trained with 60% of the data. The predicted profit is plotted against its true value both for the same training data and for the remaining test data (40% in this case). As can be seen, the model cannot be seen as a good one from a regression point of view. Nevertheless, from a classification viewpoint, this particular SVR model is able to predict profit with a positive rate of 72% and a negative rate of about 50%. These quantities are the result of blindly applying a threshold to the SVR output. But they can be appropriately tuned if needed.

Figure 4 shows a particular ROC curve obtained for a SVR model trained with 60% of the available data and computed from the remaining test data. The ROC point given by the SVR in this case is $(.67, .43)$ but as can be seen slighly better performance can be obtained for this particular model if one wants to give different importance to prediction errors in positive (profit) or negative (non profit) cases.

Figure 5 and Table I show all the results obtained for the whole range of training set sizes considered in a last experiment in which 30% of the data is kept for evaluation and the remaining 70% is used for training in an incremental way starting from 25%. As SVR have been tuned in the appropriate region of the ROC space, one can clearly see in Figure 5 an

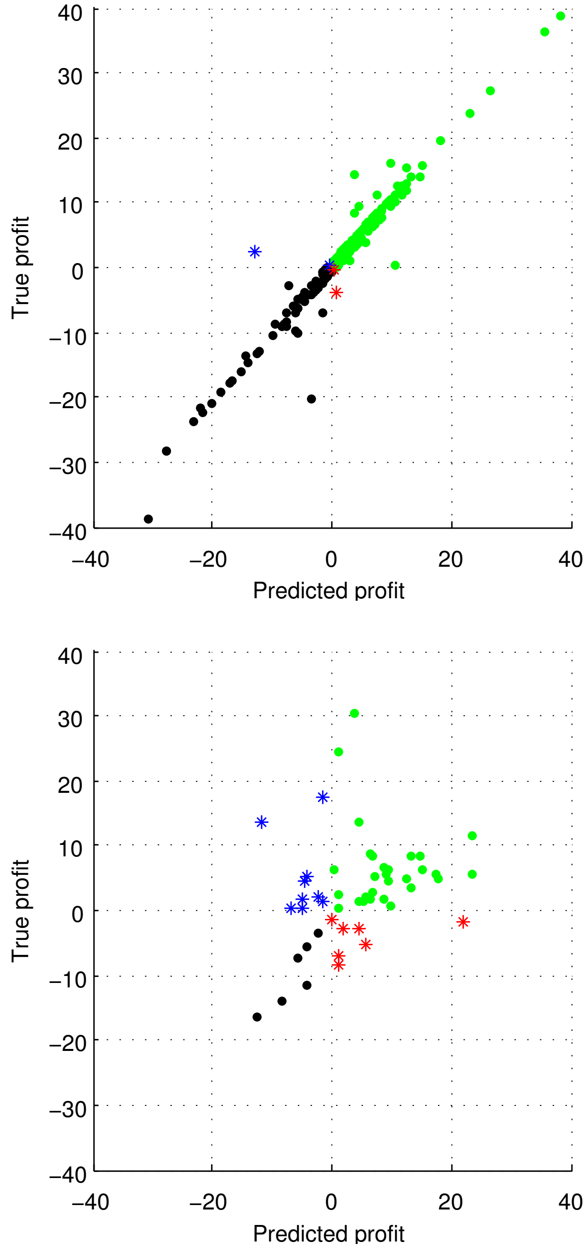

Fig. 3. Scatter plots showing predicted profits against their true values both in thousands of Euro for train (upper) and test (lower) data. Dots represent correctly predicted profit/non profit and stars represent errors.

TABLE I
AVERAGED AND BEST AUC VALUES FOR INCREASING TRAINING SET SIZES (SHOWN AS PROPORTION OF THE AVAILABLE DATA).

Training set size	0.3	0.4	0.5	0.6	0.7
Average AUC	0.577	0.577	0.586	0.602	0.600
Best AUC	0.658	0.675	0.690	0.684	0.746

increasing tendency on the positive (profit) accuracy while the negative one stays around 0.6.

To assess the learned SVR models from a general viewpoint taking into account the whole range of the ROC space, the corresponding (averaged and best) AUC values for increasing training set sizes are shown in Table I. As can be seen, both figures follow the same trend as the particular accuracy rates in Figure 5.

VI. CONCLUDING REMARKS AND CONSIDERATIONS FOR FURTHER WORK

In this work, a preliminary study about applying support vector regression to predict profit in a relatively small consulting firm using partial historical data about previous projects of the same firm has been carried out. The problem is very challenging in the particular way it has been posed. On one hand, the problem suffers from the well-known small sample size effect. On the other hand, and because of the particular data, the problem is severely biased as only data about previously completed projects by the same firm are used. Last but not least, the target data considered can be seen as very noisy as many well-known circumstances may convert some apriori profitable projects into high cost ones and the other way round. None of these details have been given to the learning algorithm in any way. Also, the temporal dimension of data (apart from including some macroeconomic indicators) has not been taken into account. Even with these known limitations, the model is able to predict profit with a moderate accuracy.

Further work is being directed towards converting these moderate results into practically acceptable ones. The presented approach will be converted into an online learning algorithm able to keep improving as new data becomes available. This is not a trivial task mainly because of the very high level of noise in the target data. Both subjective data from engineers in the firm and temporal data are to be included into the learning strategy in a near future in order to boost the final prediction power of the proposed model.

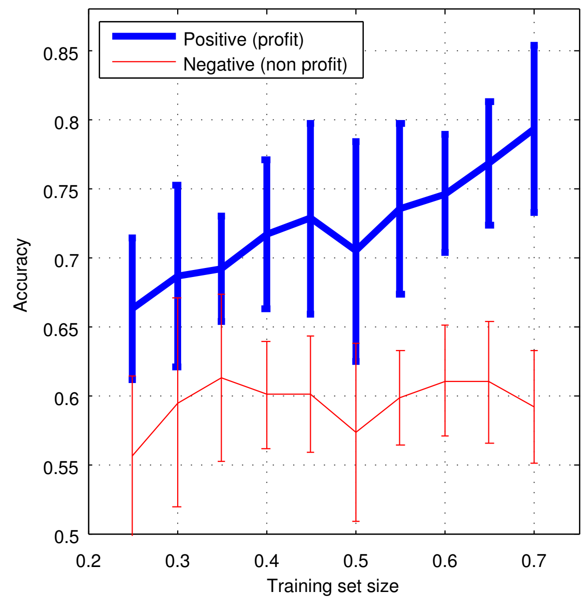

Fig. 5. Evolution of profit prediction accuracy as new training data becomes available.

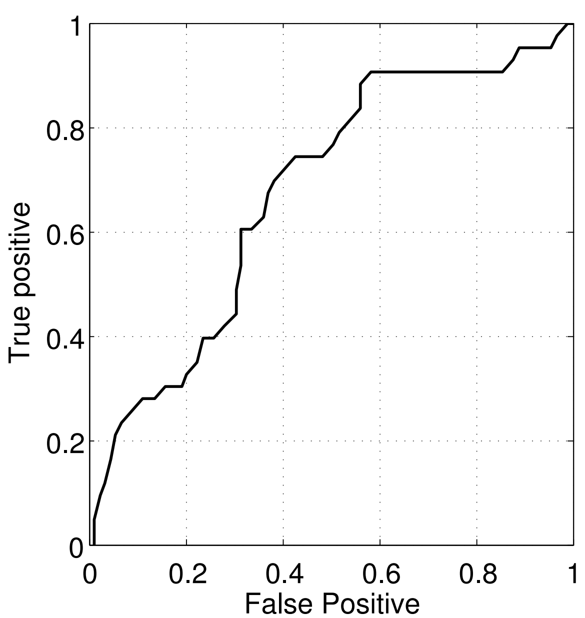

Fig. 4. ROC curve obtained for one of the SVR models obtained using 60% of the available data for training.

ACKNOWLEDGEMENT

The authors would like to thank Mr. Luis Elena for his support.

REFERENCES

[1] E. Pellicer, "Cost control in consulting engineering firms," *Journal of Management in Engineering*, vol. 21, no. 4, pp. 189–192, 2005.
[2] M. Egemen and A. Mohamed, "A framework for contractors to reach strategically correct bid/no bid and mark-up size decisions," *Building and Environment*, vol. 42, pp. 1373–1385, 2007.
[3] D. Drew and M. Skitmore, "The effect of contract type and size on competitiveness in bidding," *Construction Management and Economics*, vol. 15, pp. 469–489, 1997.
[4] L. Friedman, "A competitive-bidding strategy. operations research," vol. 4, no. 1, pp. 104–112, 1956.
[5] M. Gates, "Bidding strategies and probabilities," *ASCE Journal of the Construction Division*, vol. 93, no. CO1, pp. 75–107, 1967.

[6] I. Ahmad, "Decision-support system for modeling bid / no-bid decision problem," *Journal of Construction Engineering and Management*, vol. 116, no. 4, pp. 595–608, 1990.

[7] A. Fayek, "Competitive bidding strategy model and software system for bid preparation," *Journal of Construction Engineering and Management*, vol. 124, no. 1, pp. 1–10, 1998.

[8] H. Li and P. Love, "Combining rule-based expert systems and artificial neural networks for mark-up estimation," *Construction Management and Economics*, vol. 17, pp. 169–176, 1999.

[9] M. Wanous, A. Boussabaine, and J. Lewis, "To bid or not to bid: a parametric solution," *Construction Management and Economics*, vol. 18, pp. 457–466, 2000.

[10] D. Chua, D. Li, and W. Chan, "Case-based reasoning approach in bid decision making," *Journal of Construction Engineering and Management*, vol. 127, no. 1, pp. 35–45, 2001.

[11] P. Garzón and E. Pellicer, "Organizational efficiency of consulting engineering firms: proposal for a performance indicator," *DYNA*, 2009.

[12] B. Scholkopf and A. Smola, *Learning with Kernels*. MIT Press, 2002.

[13] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Knowledge Discovery and Data Mining*, vol. 2, no. 2, pp. 121–167, 1998.

[14] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2000.

[15] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *Machine Learning*, vol. submitted, no. http://www.hpl.hp.com/personal/Tom_Fawcett/papers/ROC101.pdf, 2004.

[16] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *International Conference on Machine Learning (ICML'97)*, 1997, pp. 179–186.