SAR and QSAR in Environmental Research,
Vol. 18, Nos. 5–6, July–September 2007, 423–441

Taylor & Francis
Taylor & Francis Group

A new descriptor selection scheme for SVM in unbalanced class problem: a case study using skin sensitisation dataset

S. LI*†‡, A. FEDOROWICZ† and M. E. ANDREW†

†Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV 26505, USA
‡Department of Statistics, West Virginia University, Morgantown, WV 26506, USA

(*Received 26 September 2006; in final form 2 February 2007*)

A novel descriptor selection scheme for Support Vector Machine (SVM) classification method has been proposed and its utility demonstrated using a skin sensitisation dataset as an example. A backward elimination procedure, guided by mean accuracy (the average of specificity and sensitivity) of a leave-one-out cross validation, is devised for the SVM. Subsets of descriptors were first selected using a sequential *t*-test filter or a Random Forest filter, before backward elimination was applied. Different kernels for SVM were compared using this descriptor selection scheme. The Radial Basis Function (RBF) kernel worked best when a sequential *t*-test filter was adopted. The highest mean accuracy, 84.9%, was obtained using SVM with 23 descriptors. The sensitivity and the specificity were as high as 93.1% and 76.6%, respectively. A linear kernel was found to be optimal when a Random Forest filter was used. The performance using 24 descriptors was comparable with a RBF kernel with a sequential *t*-test filter. As a comparison, Fisher's linear discriminant analysis (LDA) under the same descriptor selection scheme was carried out. SVM was shown to outperform the LDA.

Keywords: Support vector machine; Variable selection; Unbalanced data; Fisher's linear discriminant analysis; Skin sensitisation

1. Introduction

Many modern classification techniques have been developed and applied to QSAR modelling, for example, Neural Network, k Nearest Neighbours, Decision Tree, Random Forest and Support Vector Machine (SVM) [1], etc. In practice, SVM is usually robust to outliers and has better performance than other methods [2, 3]. As a machine learner, SVM has attracted significant research interest and has been extensively applied to classification, clustering, regression and novelty detection. In this study, we shall use SVM as a classifier. The basic SVM principle for a two-class problem will be described in the method section. For more elaborate SVM theory, refer to [1–6] and http://www.kernel-machines.org.

*Corresponding author. Email: swl4@cdc.gov or shli@stat.wvu.edu

SAR and QSAR in Environmental Research
ISSN 1062-936X print/ISSN 1029-046X online © 2007 Taylor & Francis
DOI: 10.1080/10629360701428474

In QSAR studies, usually hundreds or even thousands of descriptors are generated using computational chemistry software to characterise the structural and physiochemical nature of a molecule. However, many descriptors are noisy and irrelevant to the target activity, therefore should be excluded. Removal of extraneous descriptors will reduce the dimensionality of the feature space, alleviate overfitting, and clean the decision boundary. This improves the prediction accuracy and leads to simple, robust, and computationally efficient models for easy application and interpretation. Picking the relevant descriptors and discarding redundant ones from a large pool has been a crucial step for successful modelling. Among these classification methods, Decision Tree has an internal automatic variable selection mechanism. But for all other classifiers, descriptor selection has been a challenging issue. People mainly deal this in two ways [7]. One is the so-called filter method, which is general and independent of the target classifier, because it is used before training the classifier [8–10]. Filter methods are fast and easy to implement. The second approach is to use the wrapper method, which in contrast, is bound to a specific classifier.

Various wrapper methods for feature selection have been proposed for SVM [11–16]. These methods require modification of the SVM algorithm and thus result in different SVM implementations. Another drawback is that they are not directly related with the SVM prediction accuracy and lack of user control of the performance. One such wrapped SVM feature selection algorithm is the SVM-RFE (SVM recursive feature elimination) which uses the descriptor weight magnitudes (squared weights or absolute weights, equivalently) in the decision function as a ranking criterion of the features [12].

Another frequent problem in QSAR studies is the imbalance in class size, e.g. one class may have many more observations than the other due to data limitations. This imbalance will result in unbalanced prediction accuracies including low specificity and high sensitivity, or vice versa, for most classification methods. But in many situations, a balanced accuracy, for which both specificity and sensitivity are at least, say, 70%, may be an important need. This issue has been studied by re-sampling methods, such as, up-sampling the minority and/or down-sampling the majority, and synthetic methods that generate new synthetic data for the smaller class [12, 17–20]. However, these methods may overuse or waste some data points.

We shall propose a transparent descriptor selection method to build simple SVM models. The descriptor selection scheme will be guided by a performance criterion which can be pre-specified for different specificity and sensitivity balances in order to adapt to different applications when only unbalanced data are given. Thus, the performance balancing mechanism is incorporated into the model selection procedure.

As a case study, chemical skin sensitisation data will be used. Skin sensitisation (allergic contact dermatitis) is an immunologically mediated cutaneous reaction to a chemical substance. This response is characterised in humans by pruritis, erythema, oedema, papules, vesicles, bullae, or a combination of these. It is estimated that more than 13 million workers in the USA are potentially exposed to chemicals that can be absorbed through the skin. A worker's skin can be exposed to hazardous chemicals through direct contact with contaminated surfaces, deposition of aerosols, immersion, or splashes. The resulting contact dermatitis is one of the most common chemically induced occupational disorders, accounting for 10–15% of all occupational illnesses [21]. A variety of experimental tests have been suggested to assess the skin sensitization potential of a chemical. The recently developed and validated murine Local Lymph Node Assay (LLNA) has the unambiguous, well-defined endpoint in the

population of lymphocytes at the lymph nodes. LLNA is more objective and requires less test substance. Thus the LLNA method will be widely used for allergen chemical assessment.

Using the proposed descriptor selection scheme for the Support Vector Machine (SVM) classification method, QSAR models based on a set of molecular descriptors that predict binary skin sensitisation activity of organic chemicals will be developed using LLNA data.

2. Materials

In this study, 178 organic chemicals were selected from the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) report [22, 23]. This dataset has also been analysed by other methods [24] and will be convenient for the comparison to the results of this study. The skin sensitisation dichotomous (positive/negative) responses were based on the results of LLNA. Out of the 178 chemicals are 47 are negative and 131 positive. Chemicals included in the dataset are listed in tables 1 and 2. In this dataset, the number of positive chemicals is almost three times the negative chemicals and thus is very unbalanced. The aim of obtaining balanced sensitivity and specificity values is challenging for most classifiers. In this paper, we tried to solve this problem through the mean accuracy criterion for model selection. Another small LLNA dataset of 25 chemicals [24] (table 12) was used for validation of selected models. Originally only three chemicals were negative in this collection. However, based on recent studies vanillin and ethyl vanillin should also be considered as negative chemicals [25].

Three software packages were used in order to generate as many molecular descriptors as possible. 262 descriptors were generated by Cerius^2 (Accelrys Inc., San Diego, CA), 1204 descriptors were generated by Dragon 4.0 (http://www.talete.-mi.it) and another set of 747 descriptors were generated by Molconn-Z (eduSoft, LC, Ashland, VA). We generated all possible molecular descriptors to minimize human selection bias in selecting them. Although the so-called electrophilic hypothesis relates skin sensitisation activity with electrophilic reactivity of chemicals, it still does not explain activity or lack thereof for many chemicals, even for such simple electrophilic chemicals as bromoalkanes. Therefore, it is crucial to start with all available information as this approach might provide the best model in the end. After constant and nearly constant variables were discarded, a total number of 1314 descriptors were included in the final data set. The number of descriptors is large. The proposed descriptor selection scheme will be used to select important descriptors for SVM classification.

3. Methods

3.1 *Support vector machine principle*

For a linearly separable two-class problem of n training data points and p features (variables), let x be a vector of p components in the feature space,

Table 1. LLNA positive chemicals and activity predictions by SVMs.

No.	Chemical name	CAS number	Activity prediction			
			SVM23	SVM16	SVM38	SVM24
001	1,2-Benzisothiazol-3(2H)-one	2634-33-5	+	+	+	+
002	1,2-Dibromo-2,4-dicyanobutane	35691-65-7	+	+	+	+
003	1,4-Benzoquinone	106-51-4	−	+	+	+
004	12-Bromo-1-dodecanol	3344-77-2	+	+	+	+
005	12-Bromododecanoic acid	73367-80-3	+	+	+	+
006	1-Bromododecane	143-15-7	+	+	+	+
007	1-Bromoheptadecane	3508-00-7	+	+	+	+
008	1-Bromohexadecane	112-82-3	+	+	+	+
009	1-Bromooctadecane	112-89-0	+	+	+	+
010	1-Bromopentadecane	629-72-1	+	+	+	+
011	1-Bromotetradecane	112-71-0	+	+	+	+
012	1-Bbromotridecane	765-09-3	+	+	+	+
013	1-Bromoundecane	693-67-4	+	+	+	+
014	1-Chloromethylpyrene	1086-00-6	+	+	+	+
015	1-Chlorononane	2473-01-0	+	+	+	+
016	1-Chlorooctadecane	3386-33-2	+	+	+	+
017	1-Chlorotetradecane	2425-54-9	+	+	+	+
018	1-Ethyl-3-nitro-1-nitrosoguanidine	4245-77-6	+	+	+	+
019	1-Iodohexadecane	544-77-4	+	+	+	+
020	1-Iodohexane	638-45-9	+	+	−	−
021	1-Iodononane	4282-42-2	+	+	+	+
022	1-Iodooctadecane	629-93-6	+	+	+	+
023	1-Iodotetradecane	19218-94-1	+	+	+	+
024	1-Methyl-3-nitro-1-nitrosoguanidine	70-25-7	+	+	+	+
025	1-Propyl-3-nitro-1-nitrosoguanidine	13010-07-6	+	+	+	+
026	1-Thioglycerol	96-27-5	+	+	+	+
027	2-(*N*-Acetoxy-acetamido)fluorene	6098-44-8	−	−	+	+
028	2,3-Butanedione	431-03-8	+	+	+	+
029	2,4,5-Trichlorophenol	95-95-4	+	+	+	+
030	2,4,6-Trichloro-1,3,5-triazine	108-77-0	+	+	+	+
031	2,4-Dinitrochlorobenzene	97-00-7	+	+	+	+
032	2,4-Dinitrofluorobenzene	70-34-8	+	−	+	+
033	2,4-Dinitrothiocyanobenzene	1594-56-5	+	+	+	+
034	2-Aminophenol	95-55-6	+	+	+	+
035	2-Bromotetradecanoic acid	10520-81-7	+	+	+	+
036	2-Chloromethylfluorene	91679-67-3	+	+	+	+
037	2-Hydroxyethyl acrylate	818-61-1	+	+	+	+
038	2-Mercaptobenzothiazole	149-30-4	+	+	+	+
039	2-Methoxy-4-methylphenol	93-51-6	+	+	+	+
040	2-Methyl-4,5-trimethylene-4-isothiazolin-3-one	82633-79-2	+	+	+	+
041	3,4-Dihydrocoumarin	119-84-6	+	+	+	−
042	3,5,5-Trimethylhexanoyl chloride	36727-29-4	+	+	+	+
043	3-Acetylphenyl benzoate	139-28-6	+	+	+	+
044	3-Aminophenol	591-27-5	+	+	−	+
045	3-Methoxyphenylbenzoate	5554-24-5	+	+	+	+
046	3-Methylcatechol	488-17-5	+	+	+	+
047	3-Methylcholantrene	56-49-5	+	+	+	+
048	3-Methyleugenol	186743-26-0	+	+	+	+
049	3-Phenylenediamine	108-45-2	+	+	+	+
050	4-Allylanisole	140-67-0	+	+	+	+
051	4-Methylaminophenol sulphate	55-55-0	+	+	+	+
052	4-Methylcatechol	452-86-8	+	+	+	+
053	4-Nitrobenzyl bromide	100-11-8	+	+	+	+

(*Continued*)

Table 1. Continued.

No.	*Chemical name*	*CAS number*	*Activity prediction*			
			SVM23	*SVM16*	*SVM38*	*SVM24*
054	4-Nitrobenzyl chloride	100-14-1	+	+	+	+
055	4-Nitroso-N,N-dimethylaniline	138-89-6	+	+	+	+
056	4-Phenylenediamine	106-50-3	+	+	+	+
057	4-Vinylpyridine	100-43-6	+	+	+	+
058	5,5-Dimethyl-3-(bromomethyl) dihydro-2(3H)-furanone	154750-20-6	+	+	+	+
059	5,5-Dimethyl-3-(thiocyanatomethyl) dihydro2(3H)-furanone	154750-32-0	+	+	+	+
060	5,5-Dimethyl-3-methylenedihydro2(3H)-furanone	29043-97-8	+	+	+	+
061	5-Chloro-2-methyl-4-isothiazolin-3-one	26172-55-4	+	+	+	+
062	5-Methyleugenol	186743-25-9	+	+	+	+
063	6-Methyleugenol	186743-24-8	+	+	+	+
064	7,12-Dimethylbenz[α]anthracene	57-97-6	+	+	+	+
065	7-Bromotetradecane	74036-97-8	+	+	+	+
066	Abietic acid	514-10-3	+	−	+	−
067	Ammonium thioglycolate	5421-46-5	+	+	+	+
068	α-Naphthoflavone	604-59-1	−	+	+	+
069	Aniline	62-53-3	+	+	+	+
070	Benzopyrene	50-32-8	+	+	−	+
071	Benzoyl chloride	98-88-4	+	+	+	+
072	Benzoyl peroxide	94-36-0	+	+	+	+
073	Benzyl bromide	100-39-0	+	+	+	+
074	β-Naphthoflavone	6051-87-2	+	+	+	+
075	β-Propiolactone	57-57-8	+	+	+	+
076	Butyl glycidil ether	2426-08-6	+	+	+	+
077	Chloramine T	127-65-1	+	+	+	+
078	Chlorpromazine	50-53-3	+	+	+	+
079	Cinnamic aldehyde	104-55-2	+	+	+	−
080	Citral	5392-40-5	−	−	+	+
081	Clotrimazole	23593-75-1	+	+	+	+
082	Cocoamidopryl betaine	61789-40-0	+	+	+	+
083	Diethyl sulphate	64-67-5	+	+	+	+
084	Diethylenetriamine	111-40-0	+	+	+	+
085	Dihydroeugenol	2785-87-7	+	+	+	+
086	Dimethyl sulphostearate	99785-70-3	+	+	+	+
087	Dimethyl sulphate	77-78-1	+	+	+	+
088	Disodium 1,2-diheptanoyloxy-3,5-benzenedisulphonate	374678-48-5	+	+	+	−
089	Dodecyl methanesulphonate	51323-71-8	+	+	+	+
090	Dodecylthiosulphonate	127089-67-2	+	−	+	+
091	Ethylene glycol dimethacrylate	97-90-5	+	+	+	+
092	Ethylenediamine	107-15-3	+	+	+	+
093	Eugenol	97-53-0	+	+	+	+
094	Fluorescein isothiocyanate	25168-13-2	+	+	+	+
095	Formaldehyde	50-00-0	+	+	+	+
096	Glyoxal	107-22-2	+	+	+	+
097	Hexadecanoyl chloride	112-67-4	+	+	+	+
098	Hexyl cinnamic aldehyde	101-86-0	+	+	+	+
099	Hydroquinone	123-31-9	+	+	+	−
100	Hydroxycitronellal	107-75-5	+	+	+	+
101	Imidazolidinyl urea	39236-46-9	−	−	+	+
102	Isoeugenol	97-54-1	+	+	+	+
103	Isononanoyloxybenzene sulphonate	109363-00-0	−	+	+	+
104	Isophorone diisocyanate	4098-71-9	+	+	+	+

(*Continued*)

Table 1. Continued.

No.	Chemical name	CAS number	Activity prediction			
			SVM23	SVM16	SVM38	SVM24
105	Methyl dodecanesulphonate	2374-65-4	+	+	+	+
106	Methyl hexadecene sulphonate	54612-23-6	−	+	+	+
107	Methyl methanesulphonate	66-27-3	+	+	+	+
108	Methylene diphenyl diisocyanate	101-68-8	+	+	+	+
109	*N*,*N*-dimethyl-1,3-propanediamine	109-55-7	+	+	+	+
110	*N*-Ethyl-*N*-nitrosourea	759-73-9	+	+	+	+
111	*N*-Nitroso-*N*-methylurea	684-93-5	+	+	+	+
112	Nonanoyl chloride	764-85-2	+	+	+	+
113	Octadecanoyl chloride	112-76-5	+	+	+	+
114	Octyl gallate	1034-01-1	+	+	+	−
115	Oxazolone	1564-29-0	+	+	+	−
116	Penicillin G	61-33-6	−	−	+	+
117	Pentachlorophenol	87-86-5	+	+	+	+
118	Phenyl benzoate	93-99-2	+	+	+	+
119	Phthalic anhydride	85-44-9	+	+	+	+
120	Picryl chloride	88-88-0	+	+	+	+
121	Propyl gallate	121-79-9	+	+	−	−
122	*p*-xylene	106-42-3	+	+	+	+
123	Pyridine	110-86-1	+	+	+	+
124	Sodium 4-(2-ethylhexyloxycarboxy) benzenesulphonate	264869-77-4	+	+	+	+
125	Sodium 4-sulphophenyl acetate	46331-24-2	−	+	−	−
126	Sodium benzoyloxy-2-methoxy-5-benzenesulphonate	159783-19-4	+	+	−	+
127	Sodium benzoyloxybenzenesulphonate	56265-04-4	+	+	+	+
128	Sodium norbornanacetoxy-4-benzenesulphonate	374679-08-0	+	−	+	+
129	Tetrachlorosalicylanilide	1154-59-2	+	+	+	+
130	Tetramethyl thiuram disulphide	137-26-8	+	+	+	+
131	Trimellitic anhydride	552-30-7	+	+	−	−

and let y be a vector of class labels, $+1$ and -1, for the two classes, define a linear classifier $f(x)$ as follows:

$$f(x) = sign(w_0 + w'x), \tag{1}$$

where *sign* function takes the sign of a real number, w is a weight vector of length p, and ω_0 is the so-called bias. Then the optimal separating hyperplane, i.e. the decision boundary, will be,

$$w_0 + w'x = 0, \tag{2}$$

which is found in the middle of two parallel supporting hyperplanes,

$$w_0 + w'x = \pm 1. \tag{3}$$

A hyperplane is called the supporting hyperplane for a class if all the data points of that class lie on one side of the plane (figure 1). The distance between the supporting hyperplanes, *margin*, is maximized through quadratic programming.

When the two classes overlap, a regularization parameter C is introduced as the misclassification cost.

Table 2. LLNA negative chemicals and activit predictions by SVMs.

No.	*Chemical name*	*CAS number*	*Activity prediction*			
			SVM23	*SVM16*	*SVM38*	*SVM24*
01	1-Bromobutane	109-65-9	−	−	−	−
02	1-Bromohexane	111-25-1	−	−	−	−
03	1-Bromononane	693-58-3	−	−	+	+
04	2,4-Dichloronitrobenzene	611-06-3	−	+	+	+
05	2-Acetamidefluorene	53-96-3	+	+	+	+
06	2-Chloroethanol	107-07-3	+	−	−	−
07	2-Hydroxypropylmethacrylate	923-26-2	−	−	−	+
08	2-Nitrofluorene	607-57-8	−	−	+	+
09	3-(Benzenesulphonyloxymethyl)-5, 5-dimethyldihydro-2(3H)-furanone	154750-24-0	−	−	−	−
10	3-(Chlorobenzenesulphonyloxymethyl)-5,5-dimethyldihydro-2(3H)-furanone	154750-28-4	−	−	−	−
11	4-Aminobenzoic acid	150-13-0	−	−	−	−
12	4-Chloroaniline	106-47-8	+	+	+	+
13	4-Hydroxybenzoic acid	99-96-7	−	−	−	−
14	5,5-Dimethyl-3-(mesyloxymethyl)-dihydro-2(3H)-furanone	154750-22-8	−	−	−	−
15	5,5-Dimethyl-3-(methoxybenzenesulphonyloxymethyl) dihydro-2(3H)-furanone	154750-23-9	−	−	−	−
16	5,5-Dimethyl-3-(nitrobenzenesulphonyloxymethyl)-dihydro-2(3H)-furanone	154750-29-5	−	−	−	−
17	5,5-Dimethyl-3-(tosyloxymethyl)dihydro-2(3H)-furanone	154060-50-1	−	−	−	−
18	6-Methylcoumarin	92-48-8	−	−	+	+
19	α-Trimethylammonium-4-tolyloxy-4-benzenesulphonate	264869-81-0	+	−	−	−
20	Benzocaine	94-09-7	−	−	+	+
21	Benzoyloxy-3,5-benzene dicarboxylic acid	102059-70-1	+	+	−	−
22	Chlorobenzene	108-90-7	−	−	+	+
23	Di-2-furanylethanedione	492-94-4	+	+	−	−
24	Dimethyl isophthalate	1459-93-4	−	+	−	−
25	Ethyl methanesulphonate	62-50-0	−	−	−	−
26	Geraniol	106-24-1	+	+	+	+
27	Hexane	110-54-3	+	−	−	−
28	Hydrocortisone	50-23-7	−	−	−	−
29	Isopropanol	67-63-0	−	−	−	−
30	Kanamycin A	8063-07-8	+	−	−	−
31	Lactic acid	50-21-5	−	−	−	−
32	Methyl salicylate	119-36-8	−	−	−	−
33	N′-(4-methylcyclohexyl)-N-(2-chloroethyl)-N-nitrosourea	13909-09-6	+	+	−	−
34	Neomycin	1405-10-3	−	−	−	−
35	Octadecylmethane sulphonate	31081-59-1	+	+	+	+
36	Phenol	108-95-2	−	+	+	+
37	Phthalic acid diethyl ether	84-66-2	−	−	−	−
38	Propylene glycol	57-55-6	−	−	−	−
39	Propylparaben	94-13-3	−	−	+	−
40	Resorcinol	108-46-3	−	+	−	−
41	Salicylic acid	69-72-7	−	−	−	−
42	Streptomycin	57-92-1	−	−	−	−
43	Sulphanilamide	63-74-1	−	−	−	−
44	Sulphanilic acid	121-57-3	−	−	−	−
45	Tartaric acid	87-69-4	−	−	−	−
46	Tixocortol-21-pivalate	55560-96-8	−	−	−	−
47	Trimethylammonium-3-tolyl-ε-caprolactamide chloride	374680-04-3	−	+	−	−

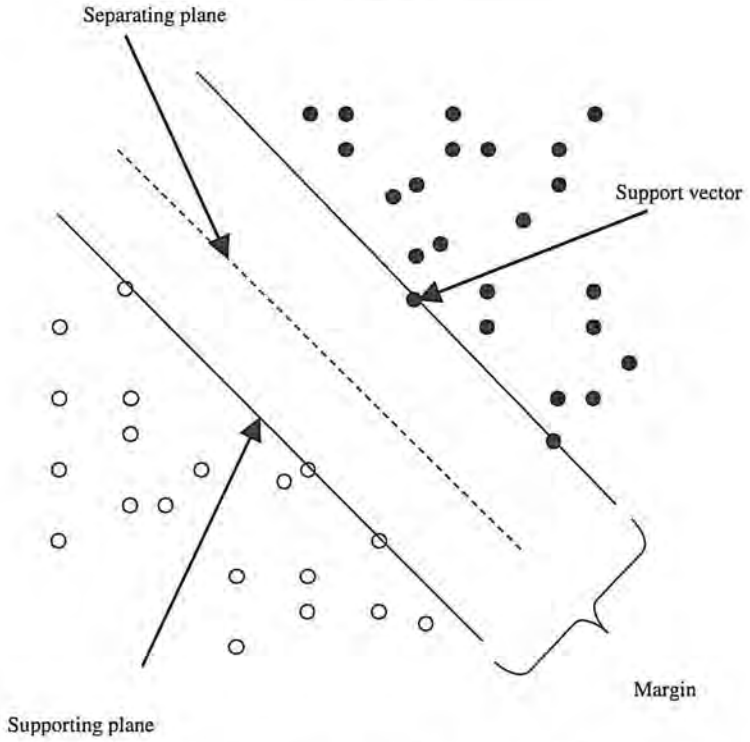


Figure 1. Support vector machine for linearly separable two classes.

A nonlinear decision function can be constructed through the *kernel* technique if the above linear classifier is not appropriate. SVM is one of the kernel methods that use kernel functions to transform raw data vectors into vectors in higher dimensional space. The theory underlying kernel techniques is sophisticated. Consider a function φ that maps a vector from the original R^p space to the higher dimensional R^d space, where $p \ll d$. In the new space, higher order and interaction terms etc. of the input variables may be added. The solution can be obtained by replacing x of the data set with $\varphi(x)$. However, this straightforward approach is of high computation cost and choosing new derived variables is difficult. The kernel method surmounts this nicely. Data participate in the computation only through the inner products in the optimization procedure. A kernel function K will give the inner product of two transformed vectors directly through untransformed vectors, i.e.:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) \equiv \varphi(\boldsymbol{x}_i)'\varphi(\boldsymbol{x}_j). \tag{4}$$

Notably, the explicit form of φ is not necessarily known and the dimensionality of φ can be infinite. There are four widely used kernels:

Linear (identity kernel, no transformation) kernel: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i'\boldsymbol{x}_j$,
Polynomial kernel: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\gamma \boldsymbol{x}_i'\boldsymbol{x}_j)^m$,
Radial Basis Function (RBF) kernel: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$,
Sigmoid kernel: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = tanh(\gamma \boldsymbol{x}_i'\boldsymbol{x}_j)$,

where γ is a tuning parameter whose value is chosen jointly with the regulation parameter C. The regulation parameter C will control the overfitting problem. Larger C will result in an overfitted decision boundary in the original vector space with smaller margin while smaller C will smooth the boundary. The cross-validation method will provide a good trade-off. The best combination of γ and C is usually located from a contour plot of the cross-validation error rate over γ and C.

3.2 *Model selection criterion*

In this study we proposed to use weighted accuracy (WA) $= \omega_1$ Sensitivity $+$ ω_2 Specificity, as the model selection criterion, i.e. the criterion for wrapped descriptor selection. For different needs, the weights can be adjusted accordingly. To get a balanced performance for the skin sensitisation study, we set equal weights, and thus the mean accuracy (MA) = (Sensitivity + Specificity)/2 was used. Other criteria, such as false negative rate and false positive rate, may also be incorporated. Interpretability of a model is difficult to measure and is not used for automatic model selection. For demonstration simplicity, we focused on sensitivity and specificity in this study. The sensitivity and specificity were obtained by Leave-One-Out Cross-Validation (LOOCV) to evaluate the different SVMs. The LOOCV method is widely used for model selection, since the error or the accuracy obtained by LOOCV is almost unbiased [5].

3.3 *Descriptor selection*

Both filter and wrapper methods were used in this work. Different filters will result in different performances for different classifiers. Unfortunately there are no established rules. In this study we will empirically compare two filters for SVM. Combining filter method and wrapper method, we will develop a new descriptor selection scheme without any changes to the SVM algorithm. First, as a quick pre-processing step before SVM modelling, we used two filters to select a subset of descriptors. One filter is a sequential t-test filter (t-filter) penalized by the correlation between descriptors:

$$\textit{Penalized } p\textit{-value} = p\textit{-value} + a\bar{r}, \tag{5}$$

where $\bar{r}$ is the mean correlation coefficient between the candidate descriptor and those already selected, and a is a constant which adjusts the correlation penalty. A new descriptor with the smallest penalized p-value is selected at each step. Through this filter, redundant descriptors can be removed from the dataset and the descriptors kept will have small correlation with each other. The other filter is based on Random Forest (RF-filter). The Random Forest [26] has two measures of descriptor importance, which provides another approach to select descriptors. We used the two-stage backward elimination algorithm proposed by Li *et al.* [27] as another filter for SVM. In this study, 100 descriptors were selected by the filters. We chose the number 100 since it is a convenient number and it is safe to keep a sufficiently large number of descriptors.

Because the filters are independent of SVM, some of the resulting descriptors may not adapt to SVM perfectly. Therefore, after the subset of descriptors is selected, extraneous descriptors need to be removed further. Though 100 is not a very large number, there is no method to find the best model from all combinations of the subset of descriptors within a reasonable amount of computing time. Different heuristic search methods may

be used, such as forward/backward/stepwise selection, genetic algorithms, evolution algorithms and simulated annealing, etc. We used the backward elimination method, which is a kind of hill-climbing or greedy search method. In the backward elimination procedure, mean accuracy is computed with one descriptor left out at each time. The descriptor corresponding to the maximal mean accuracy is deleted from the dataset step by step. The procedure is repeated until a prespecified number of descriptors are left.

3.4 *SVM kernel selection and parameter tuning*

The four kernel functions were compared and the best kernels were used in the final SVM models. As mentioned above the cost C and the kernel parameter γ need to be determined. We performed a coarse grid search over a wide range, then a fine grid search on the best small region if needed. The LOOCV mean accuracy was calculated for each pair of parameters. Based on this, the appropriate parameter set of the highest accuracy was chosen.

4. Results and discussion

4.1 *t-test filter selection*

Using the t-filter, 100 descriptors were selected first. SVM models were obtained for these descriptors with different kernels. The accuracies are tabulated in table 3. Linear, RBF and sigmoid kernels have pretty similar results, but the cubic kernel is obviously different from the others in that its specificity is low and its sensitivity is high, respectively, while its false negative and false positive rates are well balanced.

4.2 *t-test filter and backward elimination selection*

The backward elimination procedure was then applied to the above 100 descriptors to remove unnecessary descriptors for simpler models. For clarity, two figures, 2 and 3, are presented to show the mean accuracies change as the descriptors are deleted. There is an overall trend that mean accuracy increases and then decreases as the number of descriptors decreases. This trend confirms the importance of descriptor selection and the efficacy of backward elimination procedure. From figure 2, it is clear that the cubic kernel is inferior for this dataset. Interestingly, the RBF kernel based SVMs were close to linear kernel SVMs during the early stage of the process. From the point of 28 descriptors on, the former eclipsed the latter. When more descriptors were removed, the performance of the linear kernel dropped more steeply. In figure 3, the RBF kernel was compared to the sigmoid kernel. The RBF kernel is better than the sigmoid kernel

Table 3. Accuracies for different kernels using t-filter selected 100 descriptors.

Kernel	*Specificity (%)*	*Sensitivity (%)*	*False negative (%)*	*False positive (%)*	*Mean accuracy (%)*	*Total accuracy (%)*
Linear	53.2	84.0	45.6	16.7	68.6	75.9
Cubic	44.7	96.2	19.2	17.1	70.4	82.6
RBF	55.3	85.5	42.2	15.8	70.4	77.5
Sigmoid	53.2	84.7	44.5	16.5	69.0	76.4

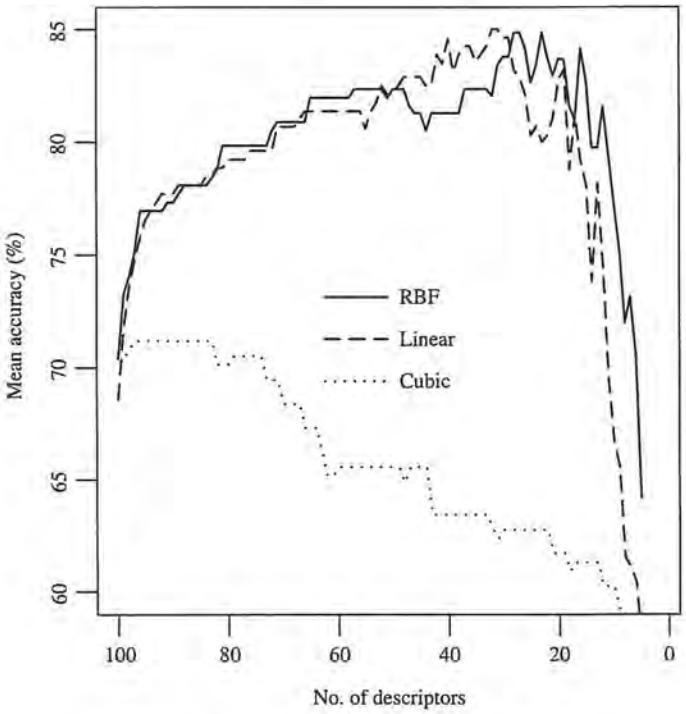


Figure 2. RBF *vs.* polynomial kernels (*t*-filter).

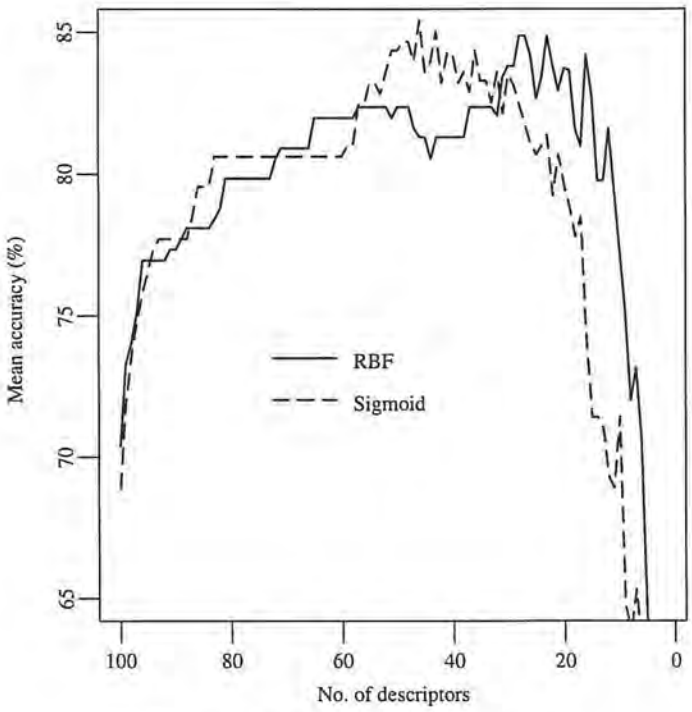


Figure 3. RBF *vs.* sigmoid kernel (*t*-filter).

when fewer than 31 descriptors are left. In summary, the RBF kernel is preferable to the linear, cubic and sigmoid kernels.

For RBF based SVM, two good models of 23 and 16 descriptors (SVM23 and SVM16) were identified from figure 3 and the comparison with the SVM of 100 descriptors is presented in table 4. The best performance was obtained for SVM23. Its sensitivity is over 90% with specificity over 70%. Mean accuracy and total accuracy are over 80%. False positive rate is less than 10% and false negative rate is around 20%. SVM16 is comparable with SVM23. Both simple SVMs performed much better than the SVM including 100 descriptors. The classification of each chemical by these two models is enclosed in table 1 and table 2, and the descriptors of the two SVMs are reported in table 5.

4.3 *Comparisons of SVM with LDA under t-test filter and backward elimination selection*

It is interesting to compare SVM with the Fisher's Linear Discriminant Analysis [28] (LDA) whose discriminant function has the same form as equation (1) but the optimal separating hyperplane is found by maximizing the between-class distance and minimizing the within-class variance simultaneously. Fisher's LDA does not assume normality, which is usually not satisfied in real data, and only uses mean and

Table 4. Accuracies of RBF kernel with different number of descriptors (*t*-filter).

No. of descriptors	*Specificity (%)*	*Sensitivity (%)*	*False negative (%)*	*False positive (%)*	*Mean accuracy (%)*	*Total accuracy (%)*
100	55.3	85.5	42.2	15.8	70.4	77.5
23	76.6	93.1	20.1	8.3	84.9	88.7
16	74.5	93.9	18.6	8.9	84.2	88.8

Table 5. Descriptors for SVM23 and SVM16.

SVM model	*Descriptor*	*Definition*
SVM23 only	n2Pag11	Vertex alpha-gamma counts
	JGI6	Topological charge indices
	MATS6e	2D autocorrelations
	R1e	GETAWAY descriptors
	Seigp	Eigenvalue-based indices
	JurS-TPSA	Jurs charged partial surface area descriptors
	Gmax	H-Bond donor/acceptor counts and EStates
SVM23 and SVM16	n2Pag22	Vertex alpha-gamma counts
	Mor11p, Mor16p, Mor22e, Mor23v	3D MoRSE descriptors
	MLOGP	Moriguchi LOGP
	JGI2	Topological charge indices
	piPC10	Walk and path counter
	MAXDN	Topological descriptors
	FDI	Geometrical descriptors
	RARS	GETAWAY descriptors
	Jurs-FNSA-3	Jurs charged partial surface area descriptors
	BIC2	Information content indices
	Shadow-nu	Surface area projections
	JX	Balaban indices
	S_ssCH2	Topological descriptors

variance information. It is a global method while SVM is a local method in the sense that the solution of the latter is determined only by the support vectors. Based on the same 100 descriptors and the backward elimination procedure, figure 4 depicts the mean accuracies of the Fisher's LDA and RBF kernel SVM for different numbers of descriptors.

In figure 4, the RBF based SVM is almost always better than Fisher's LDA. This may be explained by the benefit of kernels and margin maximization in the SVM. As seen in figure 3 and table 3, the best result for SVM is obtained when 23 descriptors are left and the SVM of 16 descriptors is still good, however the accuracy decreases sharply thereafter. The Fisher's LDA behaves a little bit differently. The best LDA model has 43 descriptors and has similar accuracies as the SVM of 23 descriptors. The LDA model of 22 descriptors is worse than the SVMs of 23 and 16 descriptors. The numeric details of these selected models are tabulated in table 6.

4.4 *RF-filter selection*

Using RF-filter, a different subset of 100 important descriptors were selected for SVM modelling. Table 7 shows the SVM classification accuracies under different kernel functions. Comparing to *t*-filter, the pattern is similar for linear, RBF and sigmoid. It is worthwhile to point out that the performance of the cubic kernel changes when switching from the *t*-filter to the RF-filter. Its specificity is higher, its sensitivity is lower and its false negative rate is much higher (changed from 19% to 56%).

4.5 *RF-filter and backward elimination selection*

Similarly, backward elimination procedure was also performed on these 100 descriptors selected by the RF-filter. The performance change is plotted in figure 5. Looking at figure 5, the cubic kernel has poor performance for any number of descriptors as was the case for the *t*-filter. A linear SVM with 38 descriptors (SVM38) gives the highest performance of the four kernels with mean accuracy 84.6%, specificity 74.5% and sensitivity 94.7%. A simpler linear SVM has 24 descriptors (SVM24). Its mean accuracy is 83% (with specificity 74.5% and sensitivity 91.6%), which is close to the RBF kernel SVM23. The results of the linear SVM38 and SVM 24 along with the one of 100 descriptors are summarized in table 8. The classifications of each chemical by SVM38 and SVM24 are reported in table 1 and table 2. The descriptors included in these two linear SVMs are listed in the table 9.

4.6 *Comparisons of SVM and LDA under Random Forest filter and backward elimination selection*

The comparison of Fisher's LDA and SVM with the linear kernel is shown in figure 6. The linear SVM is obviously better than the Fisher's LDA. The LDA works fairly well when 30–50 descriptors are included. With fewer descriptors, its performance is worse. The linear SVM does not transform the raw data, so the difference is purely due to the different optimization criteria of SVM and LDA.

4.7 *Comparison of SVM with other methods and external validation*

The training dataset was also analysed by a commercial software Derek for Windows (LHASA Ltd.) and a logistic regression model [24]. The LOOCV performance

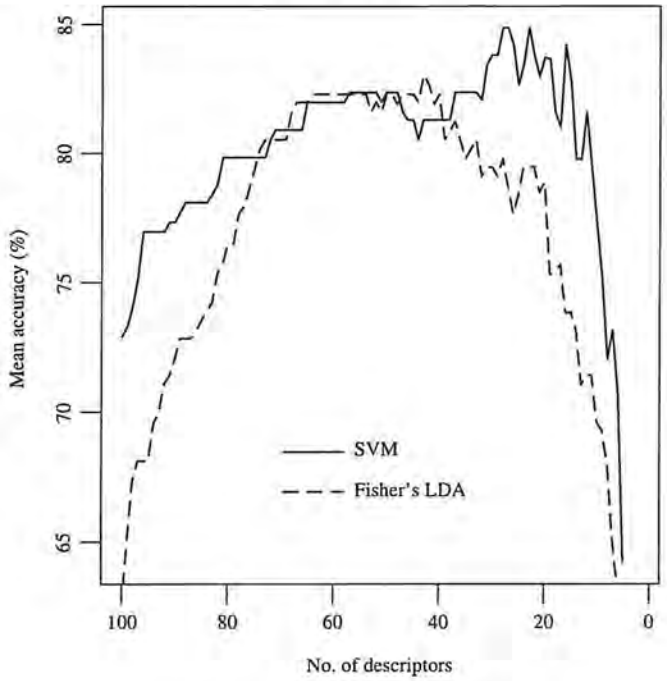


Figure 4. Fisher's LDA *vs.* RBF SVM (*t*-filter).

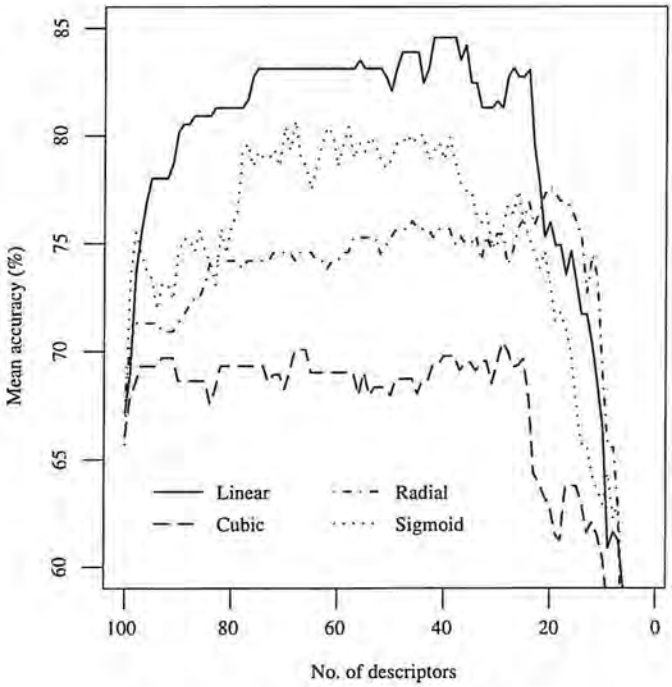


Figure 5. SVMs of different kernels (RF-filter).

Table 6. Accuracies of LDA with different numbers of descriptors.

No. of Descriptors	*Specificity (%)*	*Sensitivity (%)*	*False negative (%)*	*False positive (%)*	*Mean accuracy (%)*	*Total accuracy (%)*
100	48.9	75.6	58.2	19.5	62.3	68.5
43	74.5	91.6	23.9	9.1	83.0	87.1
22	68.1	90.8	27.4	11.2	79.5	84.8

Table 7. Accuracies for different kernels using RF-filter selected 100 descriptors.

Kernel	*Specificity (%)*	*Sensitivity (%)*	*False negative (%)*	*False positive (%)*	*Mean accuracy (%)*	*Total accuracy (%)*
Linear	51.1	83.2	47.8	17.4	67.1	74.7
Cubic	59.6	72.5	56.3	16.7	66.1	69.1
RBF	55.3	80.2	50.0	16.7	67.7	73.6
Sigmoid	48.9	82.4	50.1	18.2	65.7	73.6

Table 8. Accuracies of linear SVMs with different number of descriptors (RF-filter).

No. of descriptors	*Specificity (%)*	*Sensitivity (%)*	*False negative (%)*	*False positive (%)*	*Mean accuracy (%)*	*Total accuracy (%)*
100	51.1	83.2	47.8	17.4	67.1	74.7
38	74.5	94.7	16.7	8.8	84.6	89.3
24	74.5	91.6	23.9	9.1	83.0	87.1

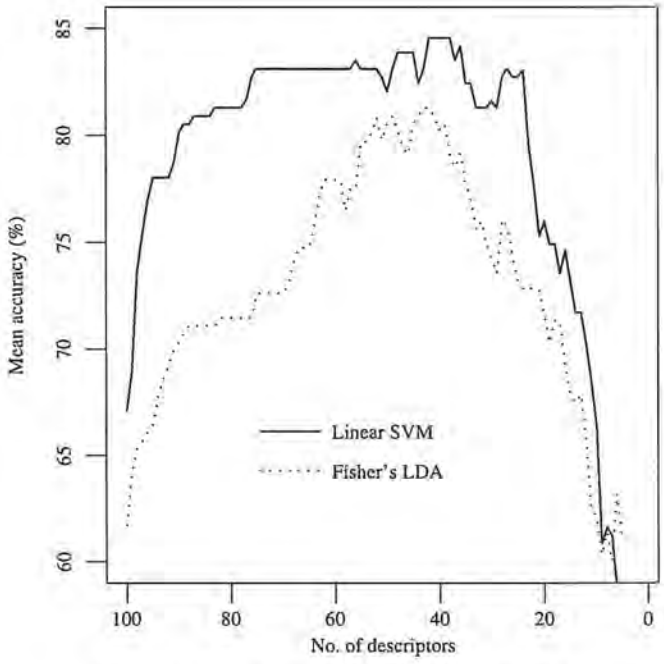


Figure 6. Fisher's LDA *vs.* linear SVM (RF-filter).

Table 9. Descriptors for SVM38 and SVM24.

SVM model	*Descriptor*	*Definition*
SVM38 only	X1sol, X2, X2v, X4Av, X4v, X5	Connectivity indices
	nDB	Double bond counter
	nCIC	Number of rings
	nCIR	Number of circuits
	nR06	Number of 6-membered rings
	Xt, Jhete, Jhetm, SPI	Topological descriptors
SVM38 and SVM24	IAC	Information indices
	RBF	Rotatable bond fraction
	RBN	Number of rotatable bonds
	nO	Oxygen atom counter
	nH	Hydrogen atom counter
	HNar, ZM1V, SMTIV, MSD, HyDp, J, Jhetp	Topological descriptors
	X0Av, X1A, X1v, X2A, X2Av, X3, X3A, X3v, X4, X5A, X5Av, X5v	Connectivity indices

Table 10. LOOCV accuracies of different methods.

Predictive models	*No. of descriptors*	*Specificity (%)*	*Sensitivity (%)*	*False negative (%)*	*False positive (%)*	*Mean accuracy (%)*	*Total accuracy (%)*
Derek for Windows	–	32.6	87.1	51.9	21.7	59.8	73.0
Logistic	5	45.7	94.7	25.0	16.7	70.2	82.0
SVM23	23	76.6	93.1	20.1	8.3	84.9	88.7
SVM16	16	74.5	93.9	18.6	8.9	84.2	88.8
SVM38	38	74.5	94.7	16.7	8.8	84.6	89.3
SVM24	24	74.5	91.6	23.9	9.1	83.0	87.1

comparison of SVM methods with these two methods are given in table 10. Derek's accuracies are the lowest among these methods. The logistic regression model is simple and uses only five descriptors. Its sensitivity is high, 94.7%, but at the cost of its specificity which is low, 45.7%. The SVM models we obtained have much higher specificity, around 75%, and similar sensitivities to the logistic model. This verifies that our SVM method based on the proposed descriptor selection scheme could improve the specificity and not sacrifice the sensitivity much for this unbalanced dataset. In terms of false negative/positive rates and mean/total accuracies, SVM models are still better than the reported logistic model and Derek for Windows results.

The three methods were also applied to the small external validation dataset. Accuracies were not calculated due to the small sample size; instead only the number of chemicals correctly classified and errors are enclosed in table 11 and the predictions of each chemical in the dataset are reported in table 12. Table 11 shows that except for the logistic model and the linear SVM38, all others models have similar performance. Based on these results it is hard to select which model is the best one and which one is the worst. The 38 descriptor SVM has more errors on negative chemicals. This may suggest that the model is too complex and overfitted on positive chemicals from the training dataset. This may lead to low efficiency in predicting any new negative chemicals. Certainly, the validation dataset is relatively small and contains only few negative chemicals. However, this reflects the phenomenon that publications are skewed towards publishing positive chemicals and leaving results for negative chemicals

Table 11. External validation accuracies of different methods.

Predictive models	*Correctly classified negative chemicals*	*Misclassified negative chemicals*	*Correctly classified positive chemicals*	*Misclassified positive chemicals*	*Total correctly classified chemicals*
Derek for Windows	2	3	14	6	16
Logistic	1	4	20	0	21
SVM23	3	2	12	8	15
SVM16	3	2	13	7	16
SVM38	1	4	19	1	20
SVM24	3	2	14	6	17

Table 12. LLNA experimental activity and QSAR predictions of the external validation data.

No.	*Chemical name*	*CAS number*	*LLNA*	*Logistic*	*Derek for Windows*	*SVM23*	*SVM16*	*SVM38*	*SVM24*
01	Acetanisole	100-06-1	−	+	+	−	−	+	+
02	Ethyl-vanillin	121-32-4	−	+	+	+	+	+	−
03	Glycerol	56-81-5	−	−	−	−	−	−	−
04	Octanoic-acid	124-07-2	−	+	−	−	−	+	−
05	Vanillin	121-33-5	−	+	+	+	+	+	+
06	2-Mercaptobenzimidazole	583-39-1	+	+	−	+	+	+	+
07	2-Methyl-4-isothiazolin-3-one	2682-20-4	+	+	+	+	+	+	+
08	3-Propylidenephthalide	17369-59-4	+	+	−	−	−	+	+
09	4-Methylhydrocinnamaldehyde	5406-12-2	+	+	+	−	−	+	+
10	4-Tert-butyl-α-methylhydrocinnamaldehyde	80-54-6	+	+	+	−	−	+	+
11	5-Methyl-2-3-hexanedione	13706-86-0	+	+	+	+	+	−	−
12	α-Amyl-cinnamaldehyde	122-40-7	+	+	+	+	+	+	+
13	α-Butyl-cinnamic-aldehyde	7492-44-6	+	+	+	+	+	+	+
14	α-Methyl-cinnamaldehyde	101-39-3	+	+	+	+	+	+	−
15	Benzylideneacetone	122-57-6	+	+	−	−	−	+	−
16	Cinnamyl-alcohol	104-54-1	+	+	−	−	−	+	−
17	Cyclamen-aldehyde	103-95-7	+	+	+	−	−	+	+
18	Diethyl-maleate	141-05-9	+	+	+	+	+	+	+
19	Diphenylcyclopropenone	886-38-4	+	+	+	+	+	+	+
20	Dodecyl-gallate	1166-52-5	+	+	+	+	+	+	−
21	Glutaraldehyde	111-30-8	+	+	+	−	+	+	−
22	Isopropyl-myristate	110-27-0	+	+	−	−	−	+	+
23	Linalool	78-70-6	+	+	−	+	+	+	+
24	Phenylacetaldehyde	122-78-1	+	+	+	+	+	+	+
25	Toluene-2-4-diisocyanate	584-84-9	+	+	+	+	+	+	+

often unpublished. In the future, to fully evaluate presented models more data including negative chemicals needs to be collected.

5. Conclusion

The proposed two-stage descriptor selection method showed the capability of choosing a small discriminatory subset from a large pool of descriptors for SVM. The SVMs obtained with this method had relatively balanced, high sensitivity and specificity. Both the nonlinear SVM of 23 descriptors through a RBF kernel under a *t*-filter,

and the linear SVM of 24 descriptors under a RF-filter provided excellent performance, with specificity >70% and sensitivity >90%. Compared with the Fisher's LDA analysis, SVM is superior in this application. In summary, the Support Vector Machine with our descriptor selection method is a competitive and effective classification approach for predicting skin sensitisation activity and can be used for other QSAR studies with a large number of descriptors and unbalanced classes.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

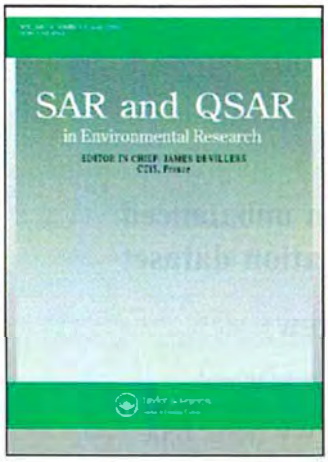
Acknowledgment

The authors thank C. Burchfiel, M. Kashon and J. E. Slaven for their valuable comments and helpful discussions that significantly improved the quality of this paper.

References

[1] C.J.C. Burges. *Data Mining Knowledge Discovery*, **2**, 2 (1998).
[2] K. Bennett, C. Campbell. SIGKDD Explorations **2**, 2 (2000).
[3] C. Cortes, V.N. Vapnik. *Mach. Learn*, **20**, 3 (1995).
[4] N. Cristianini, J. Shawe-Taylor. *An introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, Cambridge (2000).
[5] V.N. Vapnik. *Statistical Learning Theory*, John Wiley & Sons, New York (1998).
[6] A.J. Smola. Learning with Kernels, PhD thesis, Technische Universität Berlin (1998).
[7] R. Kohavi, G.H. John. *Artif. Intell.*, **97**, 1 (1997).
[8] D.C. Whitley, M.G. Ford, D.J. Livingstone. *J. Chem. Inf. Comput. Sci.*, **40**, 5 (2000).
[9] M. Hall. Correlation-based Feature Subset Selection for Machine Learning. Ph.D. thesis, Department of Computer Science, University of Waikato (1999).
[10] J. Biesiada and W. Duch. Paper presented at the Proceedings of the 4th International Conference on Computer Recognition Systems CORES'05, Springer (2005).
[11] J.-H. Lee, C.-J. Lin. Automatic model selection for support vector machines (2000). Available online at: http://www.csie.ntu.edu.tw/~cjlin/papers/modelselect.ps.gz
[12] I. Guyon, J. Weston, S. Barnhill. *Mach. Learn.*, **46**, 1 (2002).
[13] H. Fröhlich, A. Zell. Paper presented at the 2004 IEEE International Joint Conference on Neural Networks, IEEE (2004).
[14] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik. in *Advance in Neural Information Processing Systems 13*, pp. 668–674, MIT Press, Cambridge, MA (2001).
[15] Y. Grandvalet, S. Canu. *Advance in Neural Information Processing Systems 15*, MIT Press, Cambridge, MA (2003).
[16] O. Chapelle, V. Vapnik. *Advance in Neural Information Processing Systems 12*, MIT Press, Cambridge, MA (2000).
[17] J.J. Chen, C.A. Tsai, J.F. Young, R.L. Kodell. *SAR QSAR Environ. Res.*, **16**, 6 (2005).
[18] N. Japkowicz. Paper presented at the Proceedings of the 2000 International Conference on Artificial Intelligence, Las Vegas, NV (2000).
[19] N.A. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer. *J. Artif. Intell. Res. (JAIR)*, **16**, 1 (2002).
[20] A. Nickerson, N. Japkowicz, E. Milios. Paper presented at the Proceedings of Eighth International Workshop on AI and Statistics, Key West, FL (2001).
[21] C.K. Smith, S.A.M. Hotchkiss. *Allergic Contact Dermatitis*, Taylor & Francis, New York (2001).
[22] K.E. Haneke, R.R. Tice, B.L. Carson, B.H. Margolin, W.S. Stokes. *Regul. Toxicol. Pharmacol.*, **34**, 3 (2001).

[23] NIH. *NIH Publication No. 99-4494: The Murine Local Lymph Node Assay: A Test Method for Assessing the Allergic Contact Dermatitis Potential of Chemicals/Compounds*, NIH, Bethesda, MD (1999).
[24] A. Fedorowicz, H. Singh, S. Soderholm, E. Demchuck. *Chem. Res. Toxicol.*, **18**, 6 (2005).
[25] D.A. Basketter, Z.M. Wright, E.V. Warbrick, R.J. Dearman, I. Kimber, C.A. Ryan, G.F. Gerberick, I.R. White. *Contact Dermatitis*, **45**, 2 (2001).
[26] L. Breiman. *Mach. Learn.*, **45**, 1 (2001).
[27] S. Li, A. Fedorowicz, H. Singh, S.C. Soderholm. *J. Chem. Inf. Model.*, **45**, 4 (2005).
[28] W.N. Venables, B.D. Ripley. *Modern Applied Statistics with S-SPLUS*, Springer, New York (1999).



SAR and QSAR in Environmental Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100694

A new descriptor selection scheme for SVM in unbalanced class problem: a case study using skin sensitisation dataset

S. Li [ab]; A. Fedorowicz [a]; M. E. Andrew [a]
[a] Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV 26505, USA
[b] Department of Statistics, West Virginia University, Morgantown, WV 26506, USA

Online Publication Date: 01 July 2007
To cite this Article: Li, S., Fedorowicz, A. and Andrew, M. E. (2007) 'A new descriptor selection scheme for SVM in unbalanced class problem: a case study using skin sensitisation dataset', SAR and QSAR in Environmental Research, 18:5, 423 - 441
To link to this article: DOI: 10.1080/10629360701428474
URL: http://dx.doi.org/10.1080/10629360701428474