Drug Screening with Elastic-Net Multiple Kernel Learning

Kitsuchart Pasupa and Zakria Hussain and John Shawe-Taylor and Peter Willett

***Abstract*— We apply Elastic-net Multiple Kernel Learning (MKL) to the MDL Drug Data Report (MDDR) database for the problem of drug screening. We show that combining a set of kernels constructed from fingerprint descriptors, can significantly improve the accuracy of prediction, against a Support Vector Machine trained on each kernel separately. To the best of our knowledge, this is the first application of MKL to the MDDR database for drug screening.**

I. INTRODUCTION

The drug discovery process is a highly complex task, expensive and time consuming. In order to reduce the cost and time to launch a drug into the market, computational modelling is used as a tool. One important technique in this area is called "virtual screening (VS)". It aims to score molecules based on the probability of molecules being active in a database, where a set of highly scored molecules will be selected and moved on to the next stage of drug discovery.

The conventional and simplest method of VS is similarity searching. It measures the degree of similarity between query compounds and compounds in a screening database. Many methods have been introduced for various descriptors and similarity coefficients, e.g., RASCAL for graph-type [1], LINGO for string-type [2]. To improve the search performance, data fusion has also been introduced. It combines data from multiple sources in order to improve on individual results. However, the application of data fusion, is sometimes impractical because the scores computed by the different scoring functions and/or from different descriptors are usually in different units and/or signs. Moreover, some scoring functions have linear relationship to the others. Since many descriptors have been introduced in the literature, we will concentrate our work on fingerprint descriptors (binary-valued data) only. We aim to improve the performance by combining four different types of fingerprints.

Multiple Kernel Learning (MKL) was developed to help combine kernels in a principled way [3]. Given a set of kernels, researchers have developed a technique of using the well-known Support Vector Machine (SVM) algorithm style optimisation framework to pose a MKL problem. In this work, our kernels will be constructed from four different types of fingerprint descriptors, and we will show that by applying a linear kernel to these feature sets, that we can improve predictive accuracy over that obtained using each individual kernel. Furthermore, we will apply the Elastic-net version of MKL which allows one to tune the level of sparsity required for the choice of kernels, by using a weighted combination of a 1-norm and 2-norm regularisation [4].

The paper is set out as follows. In Section II we describe the background techniques of SVM and our MKL in detail. Section III describes the database used in this paper, the experimental setup and results. Finally, we conclude the paper in Section IV.

K. Pasupa is with the Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand kitsuchart@it.kmitl.ac.th

Z. Hussain and J. Shawe-Taylor are affiliated with the Department of Computer Science, University College London, London WC1E 6BT, United Kingdom z.hussain@cs.ucl.ac.uk, jst@cs.ucl.ac.uk

P. Willett is with the Information School, University of Sheffield, Sheffield S1 4DP, United Kingdom p.willett@sheffield.ac.uk

II. METHODOLOGIES

MKL algorithms [3], [5], [6] are typically modifications of the SVM algorithm [7], [8]. In the following two sections, we describe the SVM and MKL algorithms we use for the experiments.

A. Support Vector Machine

Let $\mathbf{z} = \{(\phi(x_1), y_1), \ldots, (\phi(x_m), y_m)\}$ be an m-sample of input-output pairs where inputs $\phi(x) \in \mathbb{R}^n$ are mapped using the feature mapping ϕ and $y \in \mathcal{Y} = \{-1, +1\}$. Let $w \in \mathbb{R}^n$ be an n-dimensional *weight* vector and let $\xi \in \mathbb{R}^m$ be an m-dimensional vector of slack variables. The following optimisation problem defines the *primal soft-margin SVM*:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|_2^2 + C\,\|\xi\|_1 \\ \text{w.r.t.} \quad & w \in \mathbb{R}^n, \xi \in \mathbb{R}^m, b \in \mathbb{R} \\ \text{s.t.} \quad & y_i\left(\langle w, \phi(x_i)\rangle + b\right) \geq 1 - \xi_i,\ i = 1, \ldots, m, \\ & \xi_i \geq 0,\ i = 1, \ldots, m, \end{aligned}$$

where $b \in \mathbb{R}$ is the bias, and $C \in \mathbb{R}$ is the penalty parameter. The primal problem can be converted into the dual form by using the technique of Lagrange multipliers. The *dual SVM* is written as the following optimisation:

$$\begin{aligned} \max \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,i'=1}^{m} \alpha_i \alpha_{i'} y_i y_{i'} \kappa(x_i, x_{i'}) \\ \text{w.r.t.} \quad & \alpha \in \mathbb{R}^m \\ \text{s.t.} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C,\ i = 1, \ldots, m, \end{aligned}$$

where $\alpha = (\alpha_1, \ldots, \alpha_m)^\top$ and $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle$ is called the *kernel function*.

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

B. Multiple Kernel Learning

For MKL, given a kernel function κ, we will use $\phi_\kappa(\cdot)$ to denote a feature space mapping satisfying $\kappa(x, x') = \langle \phi_\kappa(x), \phi_\kappa(x') \rangle$. In MKL, we consider a family of kernels $\mathcal{K}$ and the corresponding function class

$$\mathcal{F}_\mathcal{K} = \{ x \mapsto \langle w, \phi_\kappa(x) \rangle \mid \|w\|_2 \leq 1, \text{ for some } \kappa \in \mathcal{K} \}.$$

Let $\mathcal{K} = \{\kappa_1, \ldots, \kappa_p\}$ denote a family of kernels, where each kernel κ_j is called the jth *base* kernel. The following kernel family is formed using a convex combination of base kernels:

$$\mathcal{K}_{\text{con}}(\kappa_1, \ldots, \kappa_p) = \left\{ \kappa^\lambda = \sum_{j=1}^{p} \lambda_j \kappa_j \mid \lambda_j \geq 0, \sum_{j=1}^{p} \lambda_j = 1 \right\}.$$

This kernel family is considered *finite* dimensional. From now on, we shall use $\phi_i = \phi_{\kappa_i}$ to denote the feature space corresponding to the kernel function κ_i.

We proceed by looking for a good combination of kernels within the SVM optimisation problem, by setting up the following elastic-net MKL problem defined by [4], [9]:

$$\begin{aligned}
\min \quad & \frac{1}{2} \sum_{j=1}^{p} \frac{\mu}{\lambda_j} \|w_j\|_2^2 + \sum_{j=1}^{p} (1-\mu) \|w_j\|_2^2 + C \|\xi\|_1 \\
\text{w.r.t} \quad & w_j \in \mathbb{R}^n, \xi \in \mathbb{R}^m, b \in \mathbb{R}, \lambda \in \mathbb{R}^p, \mu \in (0, 1] \\
\text{s.t.} \quad & y_i \left(\sum_{j=1}^{p} \langle w_j, \phi_j(x_i) \rangle + b \right) \geq 1 - \xi_i, \ i = 1, \ldots, m, \\
& \sum_{j=1}^{p} \lambda_j = 1, \lambda_j \geq 0, \ j = 1, \ldots, p, \\
& \xi_i \geq 0, \ i = 1, \ldots, m,
\end{aligned}$$

where w_j is the weight vector of the j^{th} feature space ϕ_j. We will use the dual of this optimisation problem, which due to space constraints, we do not state here. We refer the interested reader to [4] for details.

III. EXPERIMENT

A. The MDL Drug Data Report Database

The MDL Drug Data Report (MDDR) database is used in this paper [10]. It is a collection of biologically relevant molecules which are compiled from journals and patent literature. Many extensive experimentations have been carried out on this dataset [11], [12], [13], [14]. There are 102,512 molecules which are represented by four types of fingerprints as follows:

1) BCI (Barnard Chemical Information) fingerprint is generated by the use of a fragment dictionary based method [15]. The dictionary gives the relationship between substructures and corresponding bits in the fingerprint. It contains 1052 bits.
2) Daylight fingerprint uses a hashing based method [16] and does not require a pre-defined fragment dictionary. Therefore, a bit is not required to have a specific assigned fragment. Moreover, a fragment may have more than one corresponding bit. The Daylight fingerprint is represented by a 2048-bit string.
3) Extended-connectivity Fingerprint (ECFP) is one of the most popular fingerprint in this application. It represents molecular structures by means of circular atom neighbourhoods [17]. ECFP_4 is used in this work. "4" indicates the maximum diameter of the circular neighbourhoods considered for each atom. It contains 1024 bits.
4) Unity fingerprint uses both structural keys and hashing based method [18]. It is represented by a 988-D binary string (60 bits of structural keys plus 928 hashed from compounds).

We are searching for 11 activity classes (I1-I11) which have been commonly used in previous research e.g. [11], [12], [13], together with the nine most diverse activity classes (D1-D9) in the MDDR database [14], as shown in Table I. The diversity can be identified by mean self-similarity values which, in this work, are calculated by a Jaccard/Tanimoto coefficient (J/T) with ECFP_4 fingerprint as shown in Table I. These values are used to indicate how similar or dissimilar the data points are to each other. They can provide a degree of homogeneity for each activity class which reflects the degree of difficulty for each screening task. If the degree of homogeneity of an activity class is high, there is a high chance of retrieving active molecules. The number of balanced training samples (50% active and 50% inactive) for each activity class (nTr) is set to approximately 20% of the number of active molecules in each activity class ($nActive$) in the database.

B. Experimental Setup

As mentioned previously, the MDDR database is widely used. Hert and his colleagues showed that the most effective approach for this dataset is to use similarity score fusion. Moreover, binary kernel discrimination (BKD) is the current state-of-the-art algorithm for this dataset [11]. We compare the elastic-net MKL with the similarity score fusion approach, namely the SUM method as our baseline. It scores each molecule by using an average similarity value calculated by a J/T with four different types of fingerprint. However, this may not be a fair comparison because in this approach, the similarity score uses only active molecules in the training set. In contrast, we consider inactive molecules in the training set too. In this work, we only investigate the linear case. Therefore, we cannot directly compare MKL with BKD. BKD makes a use of the binomial kernel which is a nonlinear function based on a radial basis function (RBF) [7]. In this work, we compare the elastic-net MKL to individual SVMs trained with linear kernels instead. It should be noted that the BKD algorithm has strong connections with kernel logistic regression [12] which also has a relationship to SVM. This leads to a similar level of performance of these algorithms in both the linear and nonlinear cases.

The experiments were run 10 times with different random data splits. Five-fold cross validation is used as a tool to tune parameters for each method e.g. C parameter in SVM.

TABLE I

THE 11 ACTIVITY CLASSES AND THE NINE MOST DIVERSE ACTIVITY CLASSES IN MDDR DATABASE.

Index	Activity Class	$nActive$	nTr	Mean Self-similarity
I1	Renin Inhibitors	1130	226	0.338±0.091
I2	Angiotensin II AT1 Antagonists	943	190	0.270±0.093
I3	HIV Protease Inhibitor	750	150	0.227±0.092
I4	Thrombin Inhibitor	803	162	0.212±0.090
I5	Substance P Antagonists	1246	250	0.180±0.080
I6	5HT3 Antagonists	752	150	0.176±0.089
I7	D2 Antagonists	395	80	0.175±0.093
I8	5HT1A Agonists	827	166	0.167±0.086
I9	5HT Reuptake Inhibitors	359	72	0.155±0.095
I10	Protein Kinase C Inhibitor	453	92	0.143±0.102
I11	Cyclo-oxygenase Inhibitor	636	128	0.132±0.078
D1	Dopamine beta-Hydroxylase Inhibitors	94	20	0.173±0.143
D2	Phospholipase A2 Inhibitors	704	142	0.153±0.085
D3	Aldose Reductase Inhibitors	882	176	0.148±0.079
D4	Aromatase Inhibitors	513	104	0.141±0.104
D5	Lipoxygenase Inhibitors	2555	512	0.135±0.058
D6	Reverse Transcriptase Inhibitors	519	104	0.134±0.084
D7	Muscarinic (M1) Agonists	848	170	0.132±0.078
D8	NMDA Receptor Antagonists	1311	262	0.125±0.069
D9	Nitric Oxide Synthase Inhibitors	377	76	0.124±0.092

The tuned parameters are selected on the basis of the sum of active rank position, e.g., if all N_A active compounds are ranked in the first N_A positions, the rank sum is minimal. For the elastic-net MKL, we investigated $\mu = \{1, 0.5, 0\}$.

C. Experiment Results

The experimental results are shown in Table II. The 20 activity classes are ranked according to the ratio of the mean self-similarity of actives and the mean similarity value between actives and inactives in the database (hereafter, RMS) by a J/T with ECFP_4 fingerprint. The percentage of the maximum number of active compounds retrieved in the top 5% of the database is reported together with the number of retained samples in each method. Fusion by similarity scores for only actives calculated from four fingerprints achieves 55.79% accuracy on average across 20 activity classes and 10 runs. Using the SVM on each fingerprint can perform better than fusing the scores on average, moreover, it can deliver sparser solutions. This is important because speed of recall is important in a VS task. Again, the performance of the task can be improved by combining four different types of fingerprints using MKL. It can be seen that elastic-net MKL with $\mu = 0$ is the best performing among other methods, followed by μ equal to 0.5 and 1, respectively. We tested the significance level of the difference between the means of two independent samples by the t-test. Elastic-net MKL outperforms most of individual SVMs ($p < 0.001$), except in the case of elastic-net MKL with $\mu = 1$, which is better than an individual SVM with Daylight and ECFP_4 fingerprints at the significance levels $p < 0.01$ and $p < 0.05$, respectively.

When $\mu = 0$, the elastic-net MKL chooses a 2-norm regularisation, which gives rise to a weight for each kernel, but does not induce sparsity amongst the choice of kernels, as shown in Fig. 1. To encourage some sparsity, μ is set to 0.5 which leads the elastic-net MKL choosing a combination of 1-norm and 2-norm regularisation. In this instance, elastic-net MKL puts more weight on Daylight and Unity fingerprints but reduces the weight of BCI and ECFP_4. The overall picture is much the same when only 1-norm MKL regularisation is used ($\mu = 1$), creating the most sparse solution. In the case of 1-norm MKL regularisation, MKL tends to choose only two fingerprints, this leads to an increase in the number of retained samples in SVMs at 75.05%, while 2-norm MKL regularisation is at 69.67% (see Table II for details).

Fig. 2 shows that the performance of the retrieval task is good when homogeneity activity classes are performed, on the other hand, the performance drops when heterogeneity (diverse) activity classes are searched for. Moreover, the more homogeneous the activity classes, the less the number of retained samples, as shown in Fig. 3.

Elastic-net MKL displays large improvements (up to 13%) with respect to the performance of individual SVMs on accuracy for those heterogeneous activity classes but small improvements (<2%) on homogeneous activity classes as shown in Fig. 4.

IV. CONCLUSION

In this paper, we have successfully combined four types of fingerprint descriptors for the purpose of drug screening, and improved the performance significantly on heterogeneous classes, with a small improvement on homogeneous ones. It should be noted that heterogeneous activity classes are much more difficult for VS. Hence, it is a good result for MKL if it does notably better on them. Furthermore, we have proposed a procedure that allows a fast speed of recall due to the sparse nature of the solutions found. Clearly, this is advantageous for the screening process. Finally, by using the MKL algorithm we can start to understand the

TABLE II

COMPARISON OF MAXIMUM ACTIVES RETRIEVED (%) IN TOP 5% OF SAMPLE ALONG WITH THE PERCENTAGE OF RETAINED SAMPLES (BELOW)

Index	Activity Classes	RMS	Data Fusion	SVM				MKL		
				BCI	Daylight	ECFP_4	Unity	$\mu = 1$	$\mu = 0.5$	$\mu = 0$
I1	Renin Inhibitors	2.114	96.03	98.27	98.06	**99.08**	98.24	98.56	98.71	98.94
				24.91	30.58	37.12	30.53	36.11	35.27	30.49
I2	Angiotensin II AT1 Antagonists	1.733	98.20	95.80	96.64	**98.74**	96.49	97.38	97.17	97.57
				49.47	47.68	52.37	46.79	51.68	55.11	48.37
D1	Dopamine beta-Hydroxylase Inhibitors	1.679	74.05	87.98	83.33	84.64	**88.81**	87.02	87.38	87.38
				82.50	90.00	97.50	85.00	93.50	93.50	93.50
I3	HIV Protease Inhibitor	1.520	70.01	83.26	85.42	**90.22**	84.24	87.67	88.58	89.90
				52.47	59.27	64.80	52.67	65.00	63.67	61.13
I4	Thrombin Inhibitor	1.518	67.40	85.80	87.47	**93.67**	87.33	88.95	89.86	92.02
				50.37	61.36	63.58	56.98	68.21	65.62	61.73
D7	Muscarinic (M1) Agonists	1.339	56.32	87.04	85.95	84.78	85.54	87.31	88.70	**89.49**
				54.06	61.59	66.53	54.00	66.12	60.88	62.00
D4	Aromatase Inhibitors	1.403	75.23	89.05	88.00	90.52	89.07	90.02	91.76	**92.28**
				56.73	63.85	77.12	55.67	70.19	67.40	65.48
I6	5HT3 Antagonists	1.339	65.89	79.87	86.00	85.27	80.03	86.06	87.31	**89.14**
				60.60	66.73	74.27	61.27	74.40	70.60	65.00
I7	D2 Antagonists	1.309	54.45	61.35	62.42	**65.30**	58.31	62.48	64.54	65.15
				78.38	82.13	87.88	80.25	90.13	87.50	89.25
I5	Substance P Antagonists	1.358	63.42	79.63	85.90	87.38	82.08	85.47	87.15	**88.53**
				52.68	56.04	65.60	50.60	65.16	61.72	58.44
I8	5HT1A Agonists	1.321	55.46	75.26	80.08	79.10	65.28	79.57	**80.60**	79.19
				59.22	64.70	70.06	62.89	73.73	69.52	71.69
I9	5HT Reuptake Inhibitors	1.305	53.00	53.81	61.67	57.34	57.28	59.81	62.85	**63.31**
				78.06	81.94	92.08	83.47	93.47	87.08	86.67
D2	Phospholipase A2 Inhibitors	1.171	36.95	49.97	60.57	60.38	52.04	57.93	61.14	**63.32**
				72.96	78.38	80.92	72.11	83.94	81.76	74.65
I10	Protein Kinase C Inhibitor	1.199	45.68	54.18	68.53	69.07	61.77	69.39	70.54	**71.89**
				74.13	82.28	86.63	76.85	88.26	87.50	83.59
D3	Aldose Reductase Inhibitors	1.174	47.10	82.75	79.85	81.28	79.12	82.53	82.95	**84.77**
				64.49	67.90	73.64	65.51	75.63	74.03	69.83
D9	Nitric Oxide Synthase Inhibitors	1.208	40.62	72.83	72.83	72.33	**78.61**	75.10	75.52	76.52
				78.82	81.45	94.74	76.84	90.66	90.53	87.63
D6	Reverse Transcriptase Inhibitors	1.155	35.35	54.07	61.73	61.18	59.01	61.78	62.12	**65.29**
				76.06	78.37	86.35	74.90	86.83	85.96	81.44
I11	Cyclo-oxygenase Inhibitor	1.113	27.41	56.40	63.25	55.96	53.99	64.79	64.49	**66.77**
				72.42	78.75	89.30	72.34	83.67	81.48	77.97
D8	NMDA Receptor Antagonists	1.117	30.11	64.24	68.56	66.30	63.30	70.06	**73.94**	73.46
				52.82	67.48	66.45	65.99	75.46	71.60	65.84
D5	Lipoxygenase Inhibitors	1.107	23.16	57.86	65.40	55.01	58.23	64.45	67.14	**69.13**
				53.03	60.12	65.39	55.29	68.87	63.93	58.65
Average			55.79	73.47	77.08	76.88	73.94	77.82	79.12	**80.20**
				62.21	68.03	74.62	64.00	75.05	72.73	69.67

importance of some descriptors over others, based on the weightings applied to each kernel. The larger the weight, the more important that descriptor is for the prediction task at hand.

For future research directions we feel that the use of non-linear kernels with the descriptors we have used may further improve performance. Also, another direction of research would be to employ different types of descriptors (i.e., graphs, strings, etc.) and a larger number of descriptors (kernels) too. Furthermore, it has been shown that the p-norm MKL algorithms perform better than the standard 1 or 2-norm variants, so we would like to apply the p-norm MKL to the drug screening problem described in this paper.

REFERENCES

[1] J. Raymond, E. Gardiner, and P. Willett, "RASCAL: Calculation of graph similarity using maximum common edge subgraphs," *Comput. J.*, vol. 45, no. 6, pp. 631–644, 2002.

[2] D. Vidal, M. Thormann, and M. Pons, "LINGO – an efficient holographic text based method to calculate biophysical properties and intermolecular similarities," *J. Chem. Inf. Model.*, vol. 45, no. 2, pp. 386–393, 2005.

[3] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, 2004.

[4] Z. Hussain and J. Shawe-Taylor, "Metric selection prototype," PinView FP7-216529 Project Deliverable Report D3.3, January 2011. [Online]. Available: http://www.pinview.eu/deliverables.php

[5] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proceedings of the twenty-first international conference on Machine learning*. New York, NY, USA: ACM, 2004.

[6] A. Argyriou, C. A. Micchelli, and M. Pontil, "Learning convex combinations of continuously parameterized basic kernels." in *Computational Learning Theory*, ser. Lecture Notes in Computer Science, vol. 3559. Springer, 2005, pp. 338–352.

[7] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[8] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge University Press, 2004.

[9] Z. Hussain, K. Pasupa, C. J. Saunders, and J. Shawe-Taylor, "Basic metric learning," PinView FP7-216529 Project

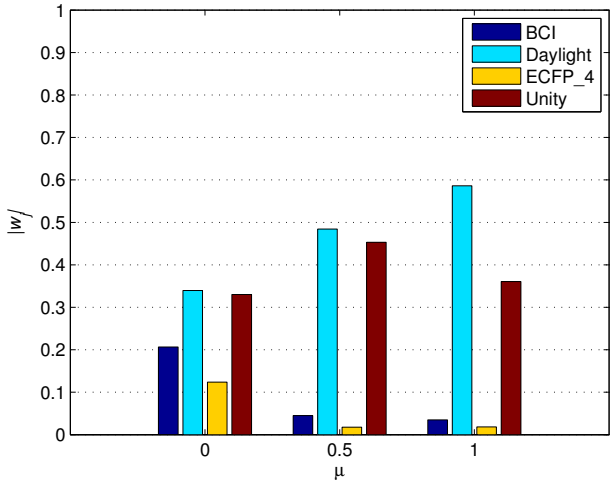

Fig. 1. Average of MKL weights (average across 10 runs, and 20 activity classes).

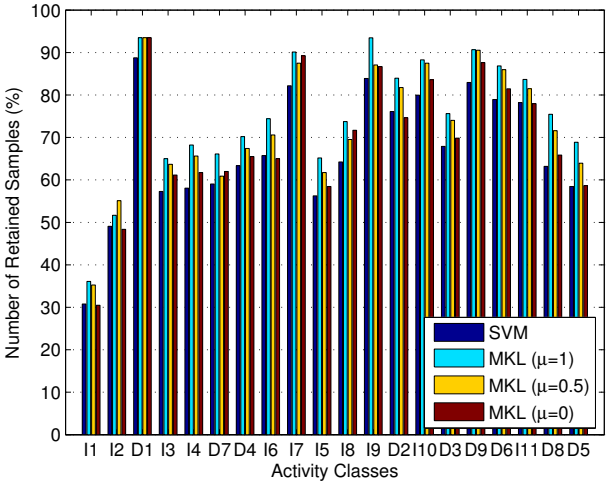

Fig. 3. Average number of retained samples by SVM and MKL for each activity class.

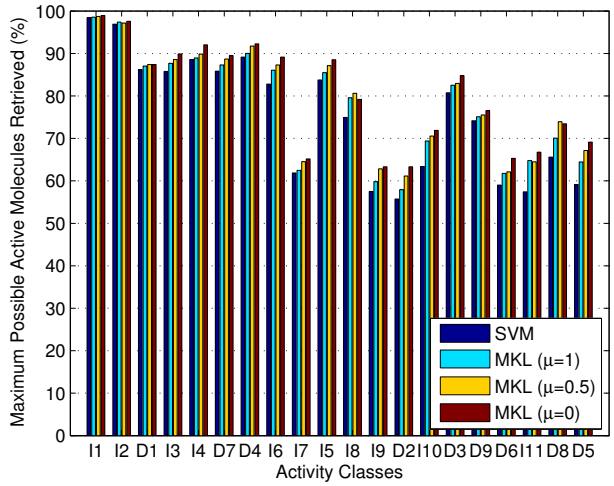

Fig. 2. Maximum possible active molecules retrieved by SVM and MKL for each activity class.

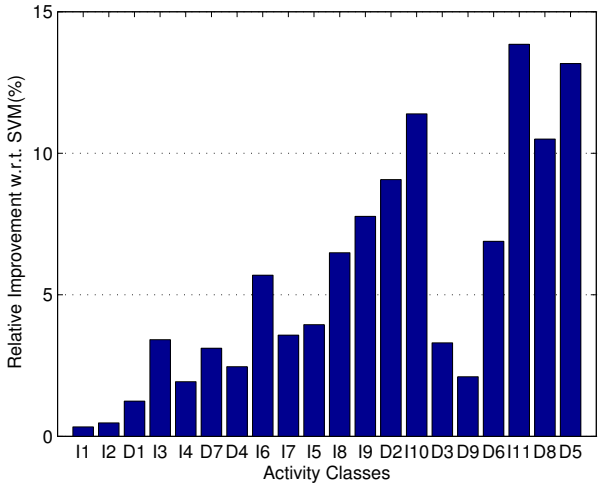

Fig. 4. Relative Improvement of MKL with respect to SVM based on active molecules retrieval rate. (average across ten runs, and 20 activity classes)

Deliverable Report D3.1, December 2008. [Online]. Available: http://www.pinview.eu/deliverables.php

[10] MDL Information Systems Inc., "The MDL drug data report database," 2006. [Online]. Available: http://www.mdli.com

[11] J. Hert, P. Willett, D. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer, "Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures," *J. Chem. Inf. Model.*, vol. 44, no. 3, pp. 1177–1185, 2004.

[12] K. Pasupa, "Data mining and decision support in pharmaceutical databases," Ph.D. dissertation, Department of Automatic Control & Systems Engineering, University of Sheffield, 2007.

[13] B. Chen, R. F. Harrison, K. Pasupa, P. Willett, D. J. Wilton, D. J. Wood, and X. Q. Lewell, "Virtual screening using binary kernel discrimination: Effect of noisy training data and the optimization of performance." *J. Chem. Inf. Model.*, vol. 46, no. 2, pp. 478–486, 2006.

[14] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer, "New methods for ligand-based virtual screening: Use of data fusion and machine learning to enhance the effectiveness of similarity searching," *J. Chem. Inf. Model.*, vol. 46, no. 2, pp. 462–470, 2006.

[15] Barnard Chemical Information Ltd., "BCI fingerprints," 2006. [Online]. Available: http://www.bci.gb.com/

[16] Daylight Chemical Information Systems Inc., "Daylight fingerprints," 2006. [Online]. Available: http://www.daylight.com/

[17] Scitegic Inc., "ECFP_4 fingerprints," 2006. [Online]. Available: http://www.scitegic.com/

[18] Tripos Inc., "Unity fingerprints," 2006. [Online]. Available: http://www.tripos.com/