BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net *Hypothesis*

Volume 7(5)

Classifying DNA repair genes by kernel-based support vector machines

Hao Jiang*, Wai-Ki Ching

Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, University of Hong Kong, Pokfulam Road, Hong Kong. Email: Hao Jiang -haohao@hkusuc.hku.hk; *Corresponding author

Received October 24, 2011; Accepted October 24, 2011; Published October 31, 2011

Abstract:
Human longevity is a complex phenotype that has a significant genetic predisposition. Like other biological processes, ageing process is governed through the regulation of signaling pathways and transcription factors. The DNA damage theory of ageing suggests that ageing is a consequence of un-repaired DNA damage accumulation. Intensive research has been carried out to elucidate the role of DNA repair systems in the ageing process. Decision Trees and Naive Bayesian Algorithm are two data-mining based classification methods for systematically analyzing data about human DNA repair genes. In this paper we develop a linearly combined kernel with Support Vector Machine (SVM) to analyze the ageing related data. The popular supervised learning algorithm enables better discrimination between ageing-related and non-ageing-related DNA repair genes. The linear combination of linear kernel and polynomial kernel of degree 3 in conjunction with SVM allows better classification accuracy in DNA repair gene data set. Compared to Decision Trees and Naive Bayesian Algorithm, SVM with the proposed kernel can achieve 65% AUC (Area Under ROC Curve) values, in contrast to 51.1% and 52.1% respectively. More importantly, we obtain 5 significant ageing-related genes selected through the training on the whole data set and they are PCNA, PARP, APEX1, MLH1 and XRCC6. Different from the two methods, we can identify another important gene PCNA in the pathways the two methods targeted, while they failed to. And two novel genes PARP, MLH1 are selected as well. The two genes might provide potential insights for biologists in ageing research. SVM is a powerful and robust classification algorithm that can yield higher predictive accuracies. The selection of proper kernel plays a more important role in fulfilling the classification task. The important genes identified not only can target critical pathways related to ageing but also detected genes that may reveal possible related ageing biomarkers.

Background:
Over the last century, human life expectancy has been steadily increasing all around the world. The major cause of mortality is no longer infectious disease. Alternatively, it can be attributed to various metabolic disorders **[1].** However, fundamental causes of the wide spread ageing process are still not clearly known and are still under investigation. Ageing research has benefited a lot from the application of genetics in the past decades. It has been argued that regulatory genes affecting multiple pathways and processes are most likely to have significant effects on longevity **[2].** Numerous ageing-related genes have already been identified **[3, 4].** DNA repair genes could be designated as a major type of genes associating with ageing process **[5–7].** Persuasive evidence has also shown that Nuclear DNA damage serves as a direct cause of ageing **[6].** In humans, most serious ageing disease is subject to DNA repair defects. To date, over 150 DNA repair genes have been identified **[8, 9].** It is not simply inherited genetics that determines who will live the longest in an energetic, disease-free state. There is no single cause of growing old, but the various mechanisms that characterize ageing are often interrelated. Scientists are identifying many interrelated pathways of ageing. This provides us with an unprecedented opportunity to gain at least partial control over this devastating process. The insulin/IGF1 signaling pathway is the best characterized pathway affecting longevity. From simpler

organisms such as C. elegans, D.mealanogaster and mouse up to humans, several studies have affirmed the positive role it played in ageing. Another critical pathway in longevity is TOR that involved in the anti-fungal effects of drug rapamycin **[1].** Base Excision Repair pathway is best established in organelles when DNA repair in organelles is characterized as a fundamental process in ageing **[10].**

However, conceptual approaches have not quite caught up with the technology. This creates an opportunity for the application of bioinformatics approaches in ageing research. Data mining, a branch of computer science **[11],** is the process of extracting patterns from large data sets. It commonly involves 4 classes of tasks: clustering, classification, regression and association rule learning. Decision Tree learning and Naive Bayesian algorithms stand as the first application of data-mining based methods for analyzing DNA repair genes **[12].**

J48 algorithm, a decision tree induction algorithm, constructs a decision tree that each interior node corresponds to one of the input variables; each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. It is simple to understand and easy to interpret by biologists. The drawback of this method comes from its generalization ability in constructing simple tree.

Naive Bayesian classifiers can be efficiently trained in a supervised learning setting through the maximization of the classification probability. The classification probability can be expressed as the product of two parts. The first part is the continuous product of empirical conditional probability of different attributes in the current data instance given that the instance belongs to some class. The second part is the empirical prior probability of the class. Over fitting in small dataset also can be detected in the use of Naive Bayesian algorithms. Two types of ageing related datasets are used to testify the robustness of the above two algorithms. One type of datasets includes only gene expression attributes. The other type of dataset involves multiple types of attributes rather than gene expression attributes. Results show that for the former type of dataset, two algorithms tend to exhibit weak performance in classification, achieving 51.1% and 52.1% AUC values respectively. More robust methods can be introduced to improve the classification accuracy for the data set. SVM constructs a hyperplane or a set of hyperplanes in a high or infinite dimensional space. The general principle for constructing a hyperplane is to minimize the generalization error of the classifier while maximizing the distance from it to the nearest data point on each side. SVMs find successful applications in many different areas. For example, in **[13],** physico-chemically weighted kernel was constructed in conjunction with SVMs for the classification of protein datasets and glycan dataset **[14].**

Recent development of kernel methods has emphasized the need to consider a combination of multiple kernels in real-world applications. An evolutionary approach was proposed for finding the optimal weights of a combined kernel used by SVMs **[15].** We apply SVM in the classification of the gene expression based ageing data. Using the linear combination of linear kernel and polynomial kernel of degree 3, better discrimination performance can be achieved. Moreover, the significant genes identified cannot only target the well-known pathway involved in ageing, but also, we identify novel genes (genes that Naive Bayesian Algorithm and Decision Tree Algorithm failed to identify). This would give potential clues for biologists for the investigation of the specific function of the selected genes.

Methodology

We employ linearly combined kernels with SVM to accomplish the task of classification for DNA repair genes. We first describe the working mechanism of SVM, and the obtaining of hyperplane after training. In a further step, the combination of kernels is introduced in order to better fit the training data. Experiments on both binary classification and significant gene selection are performed on the proposed kernel with SVMs.

Support Vector Machine

In a usual context, we try to maximize classification performance for the training data when training a classifier. Fitting to the training data and generalization ability for unknown data acts as a trade-off pair since if the classifier is too fit for the training data, generalization ability would be degraded. SVM is trained so as to maximize the generalization ability.

SVM is primarily a classifier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. The original objects are mapped using a set of mathematical functions, known as kernels. Kernels work by embedding data instances into a feature space. They have gained increasing popularity in computational biology **[24]** due to their outstanding performance in processing complicated data. The following is a flow chart of training SVM for hyperplane construction. Please see kernel function flowchart in **supplementary material.**

Combination of Kernels

One of the most important steps in SVM classification systems is the construction of appropriate kernel functions. In the case of linearly separable data, linear kernel is one of the most straightforward choices. There is no need to map data instances into a high-dimensional space. Polynomial kernel is suitable for problems where all the training data are in normalized form **[25].** As RBF kernels use the Euclidean distance, they are not robust to outliers. Real-world applications have posed a need for emphasis on the combination of kernels. Here, we propose the combination of linear kernel and polynomial kernel of degree 3 in fulfilling the task of classification of the normalized ageing data. In this context, the hyperplane can be presented as shown in **supplementary material**

Classification Accuracy

Classification accuracy is measured in terms of AUC value by 10-fold cross validation, the same as previous studies **[12].** The AUC value is the Area under of ROC curve. It measures discrimination, the ability of correctly classifying the data instances. The larger of AUC value is, the better of the predictive ability of the classification model will be. ROC analysis has been introduced in other areas like machine learning and data mining. In 10-fold cross validation, the dataset is randomly partitioned into 10 folds of approximately equal size. The 9 folds are used as training data with the remaining fold used for validation. The cross-validation process is then repeated 10 times, each of the 10 folds used exactly once as the validation data. The 10 results from the folds then can be averaged (or otherwise combined) to produce a single estimation.

Selection of Important Genes

Important genes are selected through the procedure of training for the whole dataset. In the process, a hyperplane is obtained in the meanwhile assigning different function values for each data instances. Important genes are chosen when the corresponding function value ranks in the top 5 range, i.e. [F, I] = sort (F_value); ImpG = Genelist (I (end -4, end)). Here F_value is the function value obtained through training. Genelist is the gene symbols for the whole data set. ImpG are the selected top 5 genes. As the function value to some extent measures the distance between the hyperplane and the data instance, the larger the function value is the more discriminative of the data instance from the other ones will be. The selected genes are then compared with the current biological results to see if they can target some essential pathways associated with ageing process. Furthermore, for the case of novel genes detected, biologists might find a clue for future investigation of the novel critical pathways relating to ageing.

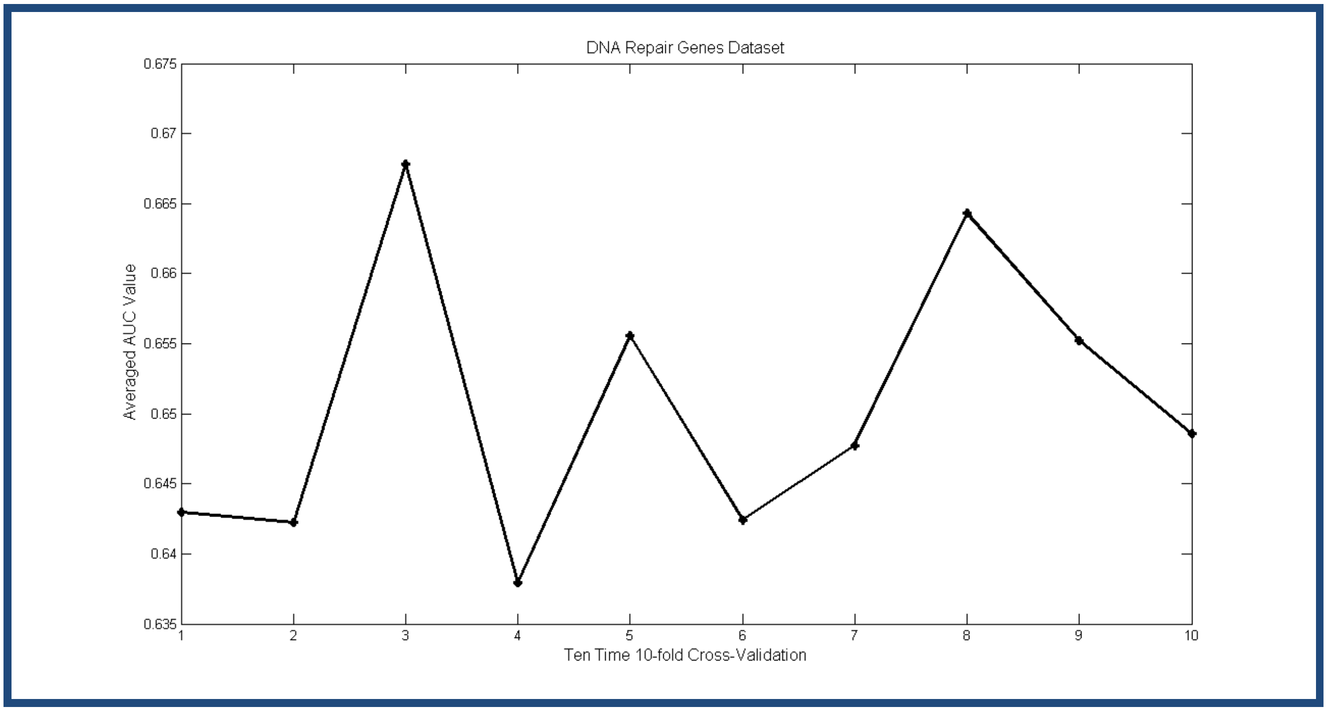


Figure 1: Performance of Combined Kernel with SVM

Results and Discussion

Materials

A set of DNA repair gene symbols was obtained from **[16]** and it consists of both positive and negative data. Ageing-related genes were included in the GenAge database, while non-ageing-related genes were not included. Attributes of the aforementioned lists of genes were obtained through Genevestigator's Anatomy tool **[17],** which is a system for analyzing gene expression and gene regulation. Each attribute correlates to an anatomical category, the value of which is the average expression level of a given gene for all probes in the corresponding anatomical category. The data set with gene expression attributes is then constructed and is shown in **Table 2 (see supplementary material).**

Classification Results

The effectiveness of our proposed method was evaluated in comparison with the J48 and Naive Bayesian Algorithm in terms of classification accuracy using the same dataset. The 10-fold cross-validation was utilized for training and testing of the data set. **Figure 1** presents the performance of the SVM classifier for the combined kernel for 10 times of 10-fold cross-validation. Here X.label stands for the time to perform 10-fold cross-validation and Y.label represents the averaged AUC value. It can be seen clearly that for the ten time 10-fold cross-validation, the AUC values can reach up to 67%. Even the lowest AUC value is around 64%, which is much higher than 51.1% and 52.1%. This is a significant improvement when compared to the previous two data-mining approaches: J48 and Naive Bayesian Algorithm. For the two methods, with the same standard of 10-fold cross-validation, they can only attain 51.1% and 52.1% respectively.

In order to get a clear idea on the superiority of our proposed combined kernel for classification in the ageing data. **Table 3 (see supplementary material)** describes in detail the performance of 10 times, 10-fold cross-validation of the model. Every row represents the performance of a 10-fold cross-validation. For instance, the first row of **Table 3 (see supplementary material)** shows the performance of first run of 10-fold cross-validation. The data set is partitioned into 10-folds. 9 out of 10-folds are selected for training the SVM model with the remaining 1 fold for testing. 0.957 is the AUC value for testing the first fold in the first run. Similarly, the procedure will repeat 9 times until all the 10 folds are used for testing. Hence, the first row illustrates that in the first run, the AUC values for the 10-folds are 0.957, 0.918, 0.837 etc. We use the Averaged AUC value in this run to ensure an unbiased measure for the performance in this run. The reason why different AUC values will be obtained for different runs of 10-fold cross-validation is that for different runs, different partitions will be generated. In this case, the models trained will be different as the data instances used for model construction will be different. But in general, the difference will be slight as can be seen from the **Table 3** (**see supplementary material)** that in each row the AUC values are around 0.65.

The construction of hyperplane is determined through training process. In the supplementary file"Svmstruct" **[18],** we report the SVMstruct used for hyperplane construction, which was included in a separate file. The support vectors, the corresponding and the bias term are included. This time there are in total 133 data instances employed for training, and the remaining 15 data instances are used for testing. In the training process, 95 data instances are selected as support vectors. Detailed information can be accessed in supplementary files. The decision function is then expressed as shown in **supplementary material.**

With the decision function, in the test of 15 data instances, we can get the function value and further determine the classes they belong to according to the sign of the function value. We present the results in a table which is attached in additional file 2 **[19].**

Selection of Important Genes

Using the whole data set in training, we can obtain a hyperplane and each instance is then assigned a score measuring the distance of the instance to the hyperplane. Five Important genes are selected according to their scores ranking and they are: PCNA, PARP1, APEX1, MLH1 and XRCC6. Compared to the genes selected by J48 and Naive Bayesian Algorithm, in the significant pathway identified, we have targeted APEX1 and XRCC6 as well. Moreover, PCNA is not included by J48 and Naive Bayes Algorithm but is detected by our method. This further validates the robustness of our proposed kernel. The novel genes not associated in the pathway are PARP1 and MLH1.

For PARP1, it is an essential component in the cellular responses to various kinds of insults to genomic DNA, including oxidative DNA damage, telomere erosion, or improper segregation of chromosomes. The PARP1proteinisakey component in repair of single-strand breaks **[20]** and it is seen as a BER modulator. PARP1 and WRN interact physically and co-operate functionally in preventing carcinogenesis in-vivo **[21]** when the WRN protein is associated with Werner's syndrome that is the one of most representative characteristics of accelerated ageing **[22].** PARP1 has also been shown to link with DNA double-strand break pathway, exhibiting various symptoms of accelerated ageing **[23].**

As for MLH1, in the "Entrez Gene summary for MLH1", it was identified as a locus frequently mutated in hereditary non-polyposis colon cancer (HNPCC). It is a human homolog of the E. coli DNA mismatch repair gene mutL, consistent with the characteristic alterations in micro satellite sequences (RER+phenotype) found in HNPCC. Alternative splicing results in multiple transcript variants encode distinct isoforms. Additional transcript variants have been described, but their natures have not been fully determined. This fact would provide some potential clues for the biologists in further investigation of the specific role MLH1 played in ageing process.

Conclusion:

We proposed a linearly combined kernel in SVM classification of DNA repair genes data set. Experiments are given to demonstrate the power of the proposed kernel when compared with the two previously proposed data-mining based approaches: J48 and Naive Bayesian Algorithm. Not only the AUC value for the classification has been improved to 10-15%, but still, the robust kernel can identify the same genes associated with the important pathways targeted by J48 and Naive Bayesian Algorithm. In a further perspective, our method detects other genes like PCNA that plays critical role in the same pathway while the two methods failed to identify. The promising perspective lies in that, we have also detected novel genes associated with ageing while the full natures of which are expecting to be explored. This would give a clue for the biologists in further investigation of the specific roles they played in ageing.

Author's contributions:

JH came up with the idea. JH and WKC designed the research. JH performed the research and analyzed the results. WKC supported the provided guidance on how to conduct the research. JH and WKC wrote the paper. Both authors read and approved the final manuscript.

Acknowledgement:

Preliminary version of the paper has been presented and published in 7th International Symposium on Bioinformatics Research and Applications. Research supported in part by GRF Grant, HKU Strategy Research Theme fund on Computational Sciences, National Natural Science Foundation of China Grant No. 10971075 and Guangdong Provincial Natural Science Grant No. 9151063101000021.

References:

[1] Riekelt H *et al. Cell.* 2010 **142**: 9

[2] Jazwinski S, *Neurobiol.* 1999 **20**: 471 [PMID: 10638520].

[3] Kenyon C, *Nature* 2010 **464**: 504

[4] De Magalhaes JP *et al. Ageing Cell* 2009 **8**: 65 [PMID: 18986374].

[5] Troen BR & Mt Sinai, *J Med.* 2003 **70**: 3 [PMID: 12516005].

[6] Best B, *Rejuvenation Research* 2009 **12**: 199 [PMID: 19594328].

[7] Hasty P *et al. Science* 2003 **299**: 1355 [PMID: 12610296].

[8] Wood R *et al. Science* 2001 **291**:1284 [PMID: 11181991]

[9] Wood R *et al. Mutation Research.* 2005 **577**:275

[10] Boesch P *et al. Biochim Biophys Acta.* 2011 **1813**:186 [PMID: 20950654].

[11] http://www.britannica.com/EBchecked/topic/1056150/data-mining.

[12] Freitas A *et al. BMC Genomics.* 2011 **12**:1 [PMID: 21226956].

[13] Jiang H *et al. Int J. Information Technology and Computer Science* 2011 **3**:1

[14] Kuboyama T *et al. Genome Informatics.* 2006 **17**:25 [PMID: 17503376].

[15] Dios L *et al. Lecture Notes in Computer Science.* 2007 **4432**:218

[16] http://sciencepark.mdanderson.org/labs/wood/DNA Repair genes.html

[17] http://www.genevestigator.com/

[18] http://hkumath.hku.hk/~wkc/papers/SVMStruct.xls

[19] http://hkumath.hku.hk/~wkc/files/Additional-file2.doc

[20] Pachkowski B *et al. Mutat. Res.* 2009 **671**:93 [PMID: 19778542].

[21] Lebel M *et al. Am J Pathol.* 2003, **162**:1559 [PMID: 12707040].

[22] Hasty P & Vijg J, *Aging Cell.* 2004, 3:55 [PMID: 15038819].

[23] Mvd V *et al. PLoS Genetics.* 2006 **2**: e192 [PMID: 17173483].

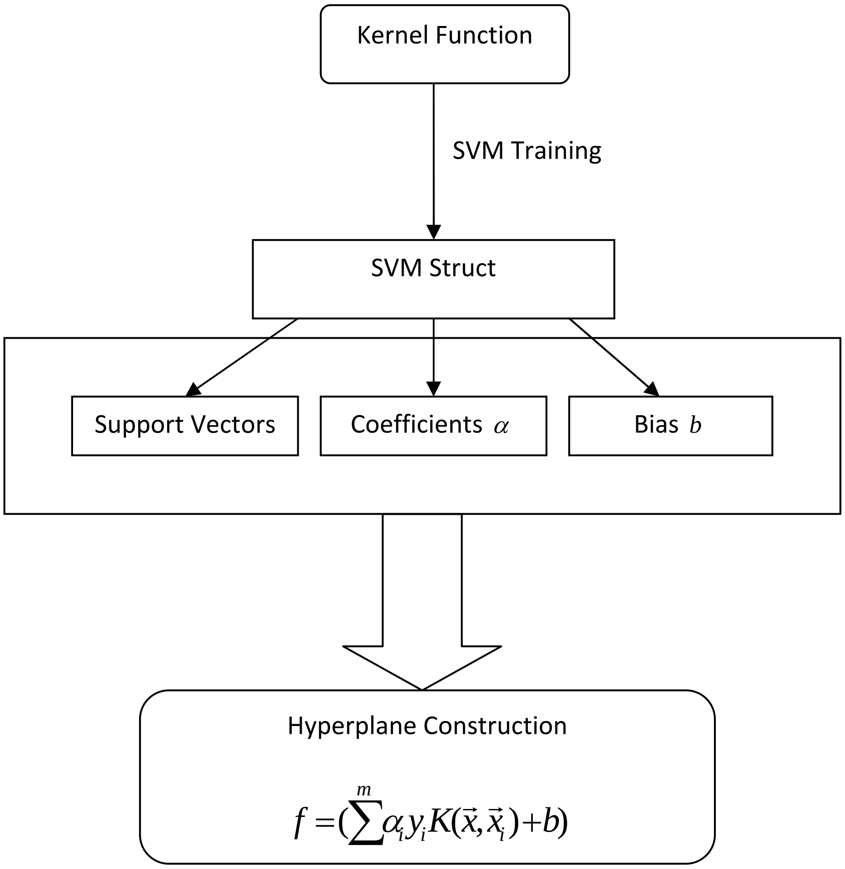
[24] Ben-Hur A & Noble WS, *Bioinformatics.* 2005 **21**: i38 [PMID: 15961482].

[25] http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html

Edited by P Kangueane

Citation: Hao Jiang & Wai-Ki Ching, Bioinformation 7(5): 257-263 (2011)

Supplementary material:



Here $\vec{x}_i, i=1,2,\cdots,m$ are the support vectors obtained from training and $\vec{\alpha}_i, i=1,2,\cdots m$ are the corresponding coefficients for the support vectors, $\vec{y}_i, i=1,2,\cdots m$ are the corresponding classes they belong to, with b being the bias part, m is the number of support vectors. The decision function is then used to evaluate the testing data to make predictions for them. If the function value of the test data instance is greater than 0, it is classified to be positive, otherwise, it is classified to be negative. Instead of being represented individually, the data are compared pairwisely and their set of pairwise similarities is represented. The matrix K contains the pairwise comparisons $K(\vec{x}_i,\vec{x}_j)$, $i, j=1,2,\cdots n$ representing the data set S, where n is the number of the data instances. Frequently used kernel functions are given in **Table 1**.

The hyperplane can be presented in the following form:

$$f=(\sum_{i=1}^{m}\alpha_i y_i[\langle\vec{x},\vec{x}_i\rangle+\langle\vec{x},\vec{x}_i\rangle^3]+b)$$

The decision function is expressed as follows:

$$f=(\sum_{i=1}^{95}\alpha_i y_i[\langle\vec{x},\vec{x}_i\rangle+\langle\vec{x},\vec{x}_i\rangle^3]+b)$$

Table 1: Kernels

Frequently Used Kernels	
Linear Kernel	$K(\tilde{x}_i, \tilde{x}_j) =< \tilde{x}_i, \tilde{x}_j >$
Polynomial Kernel	$K(\tilde{x}_i, \tilde{x}_j) = [a < \tilde{x}_i, \tilde{x}_j > +b]^c$
RBF Kernel	$K(\tilde{x}_i, \tilde{x}_j) = \exp^{\left(-\|\tilde{x}_i - \tilde{x}_j\|^2 / 2\sigma^2\right)}$

Where a, b, c, σ are kernel specific parameters.

Table 2: Dataset

DNA Repair Genes	
No. of Instances	148
Ageing-Related	33
Non-Ageing-Related	115
No. of Attributes	108

Table 3: Performance of 10-runs 10-fold cross-validation with combination of kernels in SVM

	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
1st	0.957	0.918	0.837	0.733	0.694	0.697	0.699	0.677	0.680	0.690
2nd	0.661	0.660	0.667	0.663	0.644	0.638	0.640	0.633	0.639	0.667
3rd	0.663	0.661	0.665	0.658	0.661	0.655	0.612	0.637	0.615	0.618
4th	0.620	0.614	0.610	0.638	0.630	0.614	0.611	0.605	0.635	0.632
5th	0.626	0.647	0.657	0.665	0.655	0.657	0.648	0.653	0.655	0.659
6th	0.661	0.660	0.661	0.639	0.637	0.633	0.623	0.624	0.598	0.602
7th	0.600	0.618	0.627	0.648	0.627	0.627	0.642	0.655	0.660	0.662
8th	0.675	0.678	0.680	0.677	0.674	0.671	0.686	0.689	0.684	0.685
9th	0.674	0.668	0.665	0.662	0.669	0.674	0.666	0.665	0.670	0.671
10th	0.679	0.691	0.688	0.673	0.666	0.670	0.654	0.656	0.656	0.638

Additional Files

Additional file 1 — SVMStruct

SVMStruct is included in additional file 1 under the name 'SVMStruct', it reports the SVMstruct of the trained data set of 10-fold cross-validation at one time, Support Vectors record the attribute values of the support vectors, in DG column, the corresponding support vector indices are listed. Support Coefficients report α of the corresponding support vectors, Bias is the bias part in the hyperplane.

Additional file 2— Function value and class labels

The table illustrates the function value and corresponding class labels in one run of 10-fold cross-validations. This time 15 data instances are used for testing. With the obtained decision function, function values for the 15 testing data are calculated and class labels are determined as well.