Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 94 (2016) 199 – 206

Procedia Computer Science

The 13th International Conference on Mobile Systems and Pervasive Computing (MobiSPC2016)

A new categorization numerical scheme for mobile robotic computing using odor Data-set recognition as a case

Choukri Djellali, Mehdi Adda

Mathematics, Computer Science and Engineering Dep.
University of Quebec At Rimouski
300, Allée des Ursulines, Rimouski , QC G5L 3A1, Canada

Abstract

Categorization is one of the most active research and application areas of Data Mining. In this paper, we address the problem of pattern categorization in mobile robotic computing. It is the task of automatically sorting a set of patterns into categories from a predefined set. Most categorization algorithms are sensitive to noise, architecture configuration, Bellman's curse of dimensionality, instability, and complex shapes. Hence, in the present study, a novel numerical scheme (RC) for pattern categorization which provides a good generalization ability with a small empirical error, is described. The experimental study with E-nose of six different MOX gas sensors is presented. Our evaluation method demonstrates the effectiveness and multidisciplinary applications of our approach.

Peer-review under responsibility of the Conference Program Chairs

Keywords: Machine Learning, Information Retrieval, Pattern Recognition, Categorization, Features Selection, Support Vector Machines.

1. Introduction

In Data Mining and pattern recognition, categorization is the problem of assigning each input pattern to one of a given set of categories. *Categorization can be defined as the placement of entities in groups whose members bear some similarity to each other*[1]. It is a powerful broadly applicable Data Mining technique that uses supervised learning in order to infer a complex computing function from labeled training patterns. Supervised learning entails approximating the underlying mapping between an input pattern and a desired output value (also known as label).
Nowadays, the categorization approach has been extensively used to study Mobile Communication[15], Wireless Sensor Networks[18], Mobile Web[19], Database[3], Web Services[2], robot recognition[4], video Mining[5], Information Retrieval[6], System security[8], image Mining[10] and Networking[7].

* Corresponding author. Tel.: +1-514-987-3000 ; fax: +1-514-987-8477.
E-mail address: Choukri.Djellali@uqar.ca

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2016.08.031

Most categorization algorithms are sensitive to outliers, noise, presentation order, architecture configuration and complex shapes. On one hand, the machine learning schemes deal with input patterns that are not linearly separable and the decision boundaries learned by the categorization algorithms can be complex and irregular. On the other hand, the categorization algorithms aim to converge to an optimized configuration. This state can be a local minimum of the function to be optimized (also known as loss function). This locally learning state ensures low training error and provides tight control on over-fitting but can not approximate the complex decision boundaries.
In order to avoid these limitations, we used a new categorization scheme based on kernel learning machine theory and Bootstrap aggregating scheme.
This paper is structured as follows: In Section 2, we present the current state of the art, our research questions and the problematic of categorization. The conceptual architecture of our categorization model is given in Section 3. We present in Section 4 a short evaluation with a benchmarking model for pattern categorization. Finally, a conclusion (Section 5) ends the paper with future works.

2. State of the Art, Problem and Research Questions

Categorization is one of the most important methodologies in Data Mining and it also has a central importance in pattern recognition tasks. It is considered as a separate class of supervised learning that analyzes the training patterns and produces the relevant model, i.e. representing the relationships, correlations, distribution, etc., which can be used for prediction. Supervised learning is used to infer a target function from labeled learning patterns [22].
Formally, categorization is an approximation of a target function ψ by a classifier $\tilde{\tilde{\psi}}$ which is defined as follows:

$$\begin{cases} \psi : P \times C \mapsto \{T, F\} \approx \tilde{\tilde{\psi}} : P \times C \mapsto \{T, F\} \\ \text{if } \psi(p_i, c_j) = T \rightarrow p_i \in c_j \;\; else \; p_i \notin c_j \end{cases} \tag{1}$$

The machine learning task is to select a function $\tilde{\tilde{\psi}}$ that closely approximates a target function ψ by minimizing the generalization error defined by the following formula:

$$E = \underset{\psi}{\text{Argmin}} \, (\frac{1}{n} \sum_{i=1}^{i=n} f_L(\psi(p_i), c_i)), \;\; \forall (p_i, c_i) \in S_n \tag{2}$$

Where, $P \subset \mathbb{R}^n$, $C \subset \mathbb{R}^d$ and $S_n = \{(p_1, c_1), (p_2, c_2), ...(p_n, c_d)\}$, $d \leq n.$, f_L: loss function.

Several categorization models have been suggested using machine learning as a basis for pattern recognition. Luiz M, G. Gonplves' work [9] (2000) was among the earliest efforts in which multi-feature maps are used as input to an associative memory to categorize a set of sensory patterns. In order to build the categorization model, they used a Multi-Layer Perceptron trained with a back-propagation algorithm (BPNN) and a neural network based on the Self-Organizing Map or (SOM).
In the study (Tapomayukh Bhattacharjee et al., 2013) [13] the Hidden Markov Models are used to capture the dynamic robot-environment interactions and to categorize objects. Two HMM models for categorizing trunk vs. leaf was trained. The evaluation based on cross-validation showed that the proposed algorithms yield good results.
The study of (Gonzalez-Aguirre et al., 2013) [12] presented a system to categorize small, rigid and graspable objects with limited visual sensing capabilities in a human household environment. In order to improve the categorization performance, the system used a bagging scheme based on Radial Basis Function or (RBF) kernels, MultiLayer Perceptrons or (MLP) with one hidden layer and K-nearest neighbor classifiers or (K-nn). Visual sensing from different vantage points is used to reconstructe the objects 3D mesh models. This 3D reconstruction is used for shape feature extraction.
J.R. Ruiz-Sarmiento, C. Galindo and J. Gonzalez-Jimenez (2015) [11] employed a Conditional Random Field or (CRF) model to categorize objects and rooms into robot workspace. The evaluation based on home scenes from the NYU2 Data set showed that the proposed model yields good results.
The revolution, that the mobile computing and robotic are witnessing, has led to the appearance of several categorization models. We studied a dozen or so of categorization models that originated from a variety of scientific applications,

ranging from Decision Tree models (e.g. Hunt, CART, ID3, C4.5, SLIQ, SPRINT, NBTree,..) to connectionist Neural Networks models or ANN (e.g. MLP, LVQ, RBF, BAM, Hopfield,..)[24] to probabilistic models (e.g. Naive Bayes, Multinomial, CRF, Bernoulli Multivariate,..) to Rule-based models (RIPPER, APRIORI, FP-GROWTH, CBPNARM, GSP, PrefixSpan,..) to Regression-Based models (LLSF, Logistic Regression, PLS,..) to kernel models (SV regression, KPCA, SVM[25],..) to graph models (e.g., MRF, Cyclic Pattern Kernels, Shortest Path Kernels,..) to combinatorial models (genetic algorithm or (GA)[20], simulated annealing[21], Ant colony optimization,..), etc.
Several major challenges raised in pattern categorization from pattern recognition perspective, and point out some promising research directions, particularly, the patterns inseparability, noisy patterns, representativeness, local learning, instability, etc.
Most previous models are not intended to correctly discern the structural information hidden in a collection of patterns and therefore the machine learning algorithms yield different decision boundaries. Moreover, the local learning depend heavily on the architecture configuration, Bellman's curse of dimensionality and complex shapes.
In this sense, the main goal of this work is to propose a robust categorization model that constructs complex decision boundaries and ensures low training error based on the models selection and kernel learning machine theories.

3. Architecture of our categorization system

The learning process begins with patterns presentation from a training set, as Figure 1 illustrates. In order to ensure optimum use of machine learning techniques, pretreatment of data is essential for efficient Data exploration. The features selection and normalization are the most frequently used pretreatment techniques in Data categorization. The features selection step removes noise and irrelevant variables. The second pretreatment step transforms each pattern into a normalized output vector. Thus, each vector $P_i(p_{i1}, p_{i2}, ...p_{in})$ in the learning base takes the following form as shown below:

$$\widehat{P} = \begin{pmatrix} \hat{p}_{1,1} & \cdots & \cdots & \hat{p}_{1,n} \\ \vdots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \vdots & \vdots \\ \hat{p}_{m,1} & \cdots & \cdots & \hat{p}_{m,n} \end{pmatrix}$$

Where, $0 \leq \hat{p}_{ij} \leq 1, 1 \leq i \leq m, 1 \leq j \leq n$.

The pretreatment techniques reduce the representation size and improve the execution time and learning performance. The categorization step identifies the relevant patterns in the knowledge extraction process.
In order to obtain a large margin around the decision boundaries, we used Kernel machine learning method based on Support Vector Machines or (SVM) (also support vector networks))[25]. This technique finds the maximum-margin hyperplane that represents the largest separation between a set of patterns.
Formally, the Support Vector Machine is an optimization problem that may be defined as follows:

$$\begin{cases} Max_\alpha \left\{ \displaystyle\sum_{i=1}^{i=m} \alpha_i - \frac{1}{2} \sum_{i=1}^{i=m} \sum_{j=1}^{j=m} \alpha_i \alpha_j u_i u_j K(x_i, x_j) \right. \\ \alpha_i \geq 0, \ 1 \leq i \leq m \\ \displaystyle\sum_{i=1}^{i=m} \alpha_i u_i = 0 \end{cases} \tag{3}$$

The maximum-margin hyperplane h that can accomplish this purpose can be written as $\displaystyle\sum_{i=1}^{i=m} \alpha_i^* u_i K(x, x_i) + w_0^*$.
In order to compute the dot products easily in the original space, we used hyperbolic Tangent Kernel as a Kernel function[14].
As defined in Eq. (3) and (4), the Support Vector Machine is described by an optimization Duality formulation that should find the coefficients α_i^* and w_0^* using the quadratic optimization methods[16].
The model selection step generates a diverse ensemble of classifiers by manipulating training and testing data given to a weak classifier. The purpose of this step is to select the accurate model by minimizing the bias between the

actual and the estimated accuracy. By averaging the recognition results from multiple categorization models, the generalization capability can be significantly improved. Hence, the models selection scheme is used to induce both categorization model and accuracy estimation from the same instances.

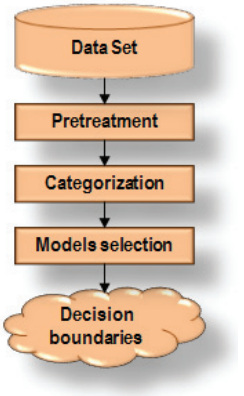

Fig. 1. Architecture of our categorization model.

4. Experimentation

4.1. Data Set

We used in our experiments the collection of a set of E-nose readings[1], which is the most widely used test collection for mobile computing. The E-nose used in this experiment contains six different MOX gas sensors: Figaro TGS-2600, TGS-2602, TGS-2611, TGS-2620, MICS-5135 and MICS-5521. Figure (2) shows the sensors readings during the E-nose aspiration to gaz pulse of Acetone analyte (10 experiments). The X-axis represents the recovery time of MOS sensors in the exposure of E-nose to the target gases and the Y-axis shows MOS gas sensor readings.

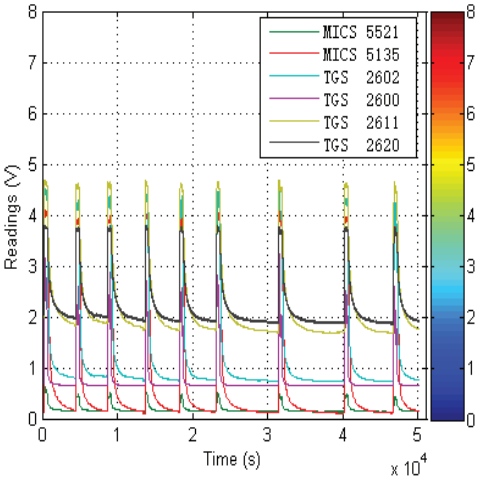

Fig. 2. The sensors readings.

[1] http://mrpt.org/robotics_datasets/

We adopted the 'ModApte' split for dividing the Data set into learning and testing sets. The details of the distribution of patterns are reported in Table (1). The number of patterns is 327137 (228996 for training and 98141 for testing). The number of patterns in each category is highly unbalanced. Thirty percent (30%) of the data are selected to test the model (no theoretical justification for this percentage). The average length of categories in terms of patterns is 46734 (32714 for training and 14020 for testing).

Table 1. The distribution of Data Sets.

Category	Learning	Test
Acetone	35076	15033
Cointreau	29637	12701
Ethanol	34521	14795
GordonGin	29519	12651
LariosGin	29707	12731
LighterGas	39407	16888
NegritaRum	31130	13341

4.2. Configuration

Our categorization scheme is developed with Java under the JEE Juno eclipse integrated development environment 64-bit and some library functions such as JDK 8u74 + Java EE, Java Matrix Package or JAMA[2], etc.

4.3. Pretreatment

After the execution of pretreatment tools, we obtained the statistical distribution as shown in Figure (3). The histogram illustrates the changes in the Data sets after pretreatment tools execution. It provides the statistical distribution of variables by comparing the proportion in which each value contributes to a total across categories. It shows the ratio of the number of noisy variables on the total number of relevant variables. The values of the series are displayed as a percentage of each class {relevant variables, noisy variables}.

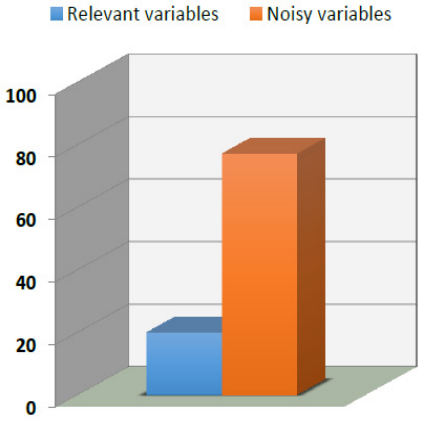

Fig. 3. The noise distribution.

[2] http://math.nist.gov/javanumerics/jama/

4.4. Categorization

Our data sets were prepared as described in section (4.1). We used a priori knowledge about the data as direct way to validate the results.
The introduction of a slight bias in our categorization model can lead to a significant reduction of its variance (a decrease of the error), and thus to improve its performance.
We used two-fold cross-validation to reduce the categorization variance. One round of cross-validation involves randomly splitting the Data set ($\mathcal{P}$) into 2 complementary folds P_L ; P_T of 2/3 of the patterns as the learning set and the remaining 1/3 as the test set. The recognition accuracy is averaged over the two rounds as defined by the following formula:

$$accuracy = \frac{1}{2}\sum_{k=1}^{2}\sum_{j=1}^{n_{Tk}}\frac{card\{recognized\ P_i\}}{card(P_{Tk})} \times 100, \quad P_i \in P_{Tk} \tag{4}$$

Learning our categorization model involves presenting input patterns in a way so that the model minimizes its loss function and improves its generalization. We used a loss (or cost) function defined as the sum over output units of the training squared difference between the desired output $d_j^{|L|}$ and the actual output $o_j^{|L|}$.

This criterion is defined as follows: $f_L = \frac{1}{n_L}\sum_{j=1}^{n_L}(e_{1j}^{|L|})^2 = \frac{1}{n_L}\sum_{j=1}^{n_L}(d_j^{|L|} - o_j^{|L|})$ (5)

$d_j^{|L|}, o_j^{|L|}, e_{1j}^{|L|}$: are respectively the desired output, actual output and signal error.

In order to select the best categorization model, we used Accuracy measure that is widely used in pattern recognition and machine learning, $Accuracy = \frac{tp + tn}{tp + fp + fn + tn}$ (6)

where, tp, tn, fp and fn represent the true positives, true negatives, false positives and false negatives.
Figure (4) shows squared error after patterns presentation. The X-axis represents the number of iterations and Y-axis shows the sum over output units of the training squared difference between the desired output and the actual error output. The performance of categorization is assessed in term of recognition accuracy during the test step. The recognition accuracy of the first model is equal to 95.87% after 53 iterations. The second model learns after 71 iterations with recognition accuracy equal to 92.79%. The first model (SVM1) significantly improves the quality of categorization. Therefore, we used this recognition model as relevant pattern discovery model to find the closely related patterns.

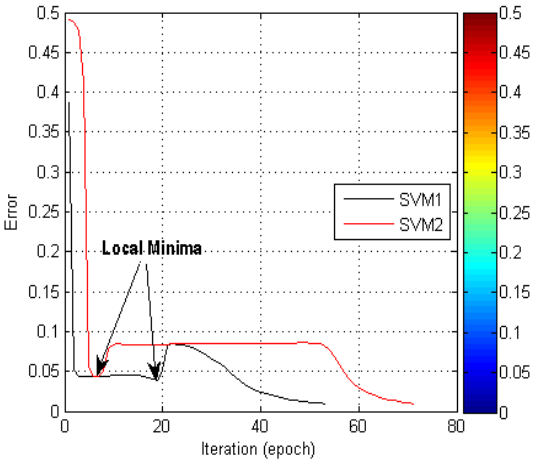

Fig. 4. The learning step.

According to the above statistical evaluation measure, our categorization model learns the complex decision boundaries with a small empirical error. It also reduces variance and avoids local minima.

4.5. Evaluation

In order to asses the performance of our proposed model, two categorization models were evaluated on the task of pattern recognition. They correspond to the neural network fuzzy ARTMAP[23] and our categorization model RC.
Fuzzy ARTMAP is a neural network architecture based on Adaptive Resonance Theory that learns to categorize patterns by fuzzy variables. Table 2 shows the Fuzzy ARTMAP architecture configuration. The ascending weights $b_{ij}^{(0)}$ are initialized by low values and backward weights $t_{ij}^{(0)}$ are initialized by the value 1. The resonance parameter ρ controls the number of neurons in the output layer. When the resonance value increases, the number of categories in the output layer also increases (the typical value of α is 0.9). The parameter α (choice parameter) takes its values in the range [0,∞[(the typical value of α is 0.001). The parameter L (uncommitted choice parameter) takes values in the interval [1,∞[. The learning rate β is placed in the interval [0,1] (the typical value of β is 0.9)[23],[17].

Table 2. fuzzy ARTMAP Architecture configuration.

Parameter	Allowable value	Typical value
L	$L > 0$	1
ρ	$0 < \rho \leq 1$	0.9
b_{ij}	$0 < b_{ij}^{(0)} < \frac{L}{L-1+N}$	0.0001
t_{ij}	$t_{ij}^{(0)} = 1$	1
α	[0,∞[	0.001
β	[0,1]	0.9

We used *precision*, *recall* and $F - measure$ indexes to validate the categorization results. These measures are widely used in pattern recognition and Data Mining.

$$precision = \frac{tp}{tp + fp} \tag{7}$$

$$recall = \frac{tp}{tp + fn} \tag{8}$$

$$F{-}measure = 2{\times}\frac{precision \times recall}{precision + recall} \tag{9}$$

Experiments (Table 3) show that our model has good performance, which provides a system of effective knowledge categorization.

Table 3. Categorization accuracy.

Model	Precision	Recall	F-measure
fuzzy ARTMAP	81.91	79.25	80.55
SC	93.87	85.75	89.62

One of the main advantages of our categorization model is its ability to directly construct the largest margin around patterns; therefore, the lowest average generalization error of the categorization. This feature was an ingredient key in the identification of complex decision boundaries.

5. Conclusion

In this paper, a categorization model dedicated to pattern recognition in mobile computing has been presented. The categorization model was tested using two-fold cross-validation, where the Data samples are used to determine the architecture and to estimate the recognition accuracy. The estimated accuracy recognition from cross-validation is not based on a selected model, but the average error of the trained models. This method optimizes the bias-variance tradeoff of the expected prediction of our categorization model.
The convergence speed of our categorization model is based on typical initializations. This initialization scheme reduces the computation time and improves the convergence speed to achieve the neighborhood vicinity of the response.

In addition, our categorization model gives an approximation function with good generalization ability.
These promising results show that our categorization model enables mobile computing to deal with complex shapes and consequently with global optimized configuration. An alternative approach for improving the recognition is to use Boosting Theory (hypothesis complexity variation). Hence, the purpose of our next work is to develop a new categorization model based on variables selection and Boosting methods.

References

1. Elin K. Jacob. Classification and categorization: a difference that makes a difference. (2004). LIBRARY TRENDS, Vol. 52, 2004, pp. 515540.
2. Jiaying He and Peipei Li and Xuegang Hu and Xindong Wu. A New Automatic Categorization Algorithm for Web Services. 2010 IEEE International Conference on Granular Computing (GrC), 2010, p. 281-304.
3. X. Meng and J. Sun and Chunxiao Liu. Context-sensitive automatic categorization of web database query results. 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). 2010, p. 1525-1529
4. D. Gonzalez-Aguirre and J. Hoch and S. Rhl and T. Asfour and E. Bayro-Corrochano and R. Dillmann.Towards shape-based visual object categorization for humanoid robots. 2011 IEEE International Conference on Robotics and Automation (ICRA). 2011, p. 5226-5232.
5. Zhang, Yimeng, et al. Object color categorization in surveillance videos. , 2011 18th IEEE International Conference on Image Processing (ICIP). IEEE, 2011, p. 2913-2916.
6. De Rooij, Ork and Worring, Marcel Active bucket categorization for high recall video retrieval. IEEE Transactions 15, 2013, p. 898-907.
7. Nguyen, Thuy TT, and Grenville Armitage. A survey of techniques for internet traffic classification using machine learning. Communications Surveys Tutorials, IEEE 10.4, 2008, p. 56-76.
8. Zabidi, Muhammad Najmi Ahmad, Mohd Aizaini Maarof, and Anazida Zainal. Ensemble based categorization and adaptive model for malware detection. Information Assurance and Security (IAS), 2011 7th International Conference on. IEEE, 2011, p. 80-85.
9. Goncalves, Luiz MG, et al. Neural mechanisms for learning of attention control and pattern categorization as basis for robot cognition. , 2000.(IROS 2000). Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vol. 1. IEEE, 2000, p. 70-75.
10. Chen, Yixin and Wang, James Z. Image categorization by learning and reasoning with regions. The Journal of Machine Learning Research. 2004 p. 913-939
11. Ruiz-Sarmiento, J. R., C. Galindo, and J. Gonzlez-Jimnez. Joint categorization of objects and rooms for mobile robots., 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2015 p. 2523-2528.
12. Gonzalez-Aguirre, D., et al. "Towards shape-based visual object categorization for humanoid robots." Robotics and Automation (ICRA), 2011 IEEE International Conference on. IEEE, 2011.
13. Bhattacharjee, Tapomayukh, et al. Rapid categorization of object properties from incidental contact with a tactile sensing robot arm. Humanoid Robots (Humanoids), 2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids). IEEE, 2013, p. 219-226.
14. Venu, Nookala, and B. Anuradha. "Integration of hyperbolic tangent and Gaussian kernels for Fuzzy C-Means algorithm with spatial information for MRI segmentation. The 2013 Fifth International Conference onAdvanced Computing (ICoAC) , . IEEE, 2013. p. 280-285.
15. Kumar, Raghavendra Phani, Malleswara Rao, and Dsvgk Kaladhar. Data Categorization and Noise Analysis in Mobile Communication Using Machine Learning Algorithms. Wireless Sensor Network 4.4 (2012):2, p. 113.
16. Huang, Guang-Bin and Ding, Xiaojian and Zhou, Hongming. Optimization method based extreme learning machine for classification j. Neurocomputing v74, 2010, p. 155-163.
17. Djellali, Choukri. A new conceptual model for dynamic text clustering Using unstructured text as a case. Proceedings of the 2014 International conference on Computer Science Software Engineering. ACM, 2014 p. 13.
18. Mohammad Abu Alsheikh1, Shaowei Lin , Dusit Niyato and Hwee-Pink Tan. Machine learning in wireless sensor networks: Algorithms, strategies, and applications." Communications Surveys Tutorials, IEEE 16.4 (2014), 4, p. 1996-2018.
19. Bouidghaghen, Ourdia, Lynda Tamine, and Mohand Boughanem. Personalizing mobile web search for location sensitive queries. The 12th IEEE International Conference on Mobile Data Management (MDM), . Vol. 1. IEEE, 2011 p. 110-118.
20. Del Castillo, M. Dolores, and Jos Ignacio Serrano. A multistrategy approach for digital text categorization from imbalanced documents. ACM SIGKDD Explorations Newsletter 6.1 (2004): p. 70-79.
21. Tian, Tian Siva, Rand R. Wilcox, and Gareth M. James. Data reduction in classification: A simulated annealing based projection method. Statistical Analysis and Data Mining 3.5 (2010) p. 319-331.
22. Cunningham, Pádraig and Cord, Matthieu and Delany, Sarah Jane. Supervised learning. Machine learning techniques for multimedia. Springer, (2008) p. 21-49.
23. Connolly, J-F., Eric Granger, and Robert Sabourin. On the correlation between genotype and classifier diversity. The 21st International Conference on Pattern Recognition (ICPR), . IEEE, 2012 p. 1068-1071.
24. Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification. John Wiley Sons, 2012.
25. Hofmann, Thomas, Bernhard Schlkopf, and Alexander J. Smola. Kernel methods in machine learning. The JSTOR, 2008, p. 1171-1220.