Individual projects

A. Approximation of an input data item by a linear mixture of SOM models

Teuvo Kohonen

The purpose of this work was to extend the use of the SOM by showing that instead of a single 'winner' model, one can approximate the input data item more accurately by means of a set of *several models* that *together* define the input data item more accurately. It shall be emphasized that we do not mean 'k winners' that are rank-ordered according to their matching. Instead, the input data item is approximated by an *optimized linear mixture of the models, using a nonlinear constraint*, which will be shown to provide an improved description of it.

Consider the n-dimensional SOM models $\mathbf{m}_i, i = 1, 2, \ldots, p$, where p is the number of nodes in the SOM. Their general linear mixture is written as

$$k_1\mathbf{m}_1 + k_2\mathbf{m}_2 + \ldots + k_p\mathbf{m}_p = \mathbf{M}\,\mathbf{k} \ , \tag{I.1}$$

where the k_i are scalar-valued weighting coefficients, $\mathbf{k}$ is the p-dimensional column vector formed of them, and $\mathbf{M}$ is the matrix with the $\mathbf{m}_i$ as its columns. Now $\mathbf{M}\,\mathbf{k}$ shall be the *estimate* of some input vector $\mathbf{x}$. The vectorial fitting error is then

$$\mathbf{e} = \mathbf{M}\,\mathbf{k} - \mathbf{x} \ . \tag{I.2}$$

Our aim is to minimize the norm of $\mathbf{e}$ in the sense of least squares. However, a special constraint must then be taken into account.

Fitting with the nonnegativity constraint

Much attention has recently been paid to least-squares problems where the fitting coefficients are constrained to *nonnegative values*. Such a constraint is natural, when the *negatives* of the items have no meaning, for instance, when the input item consists of statistical indicators that can have only nonnegative values, or is a weighted word histogram of a document. In these cases at least, the constraint contains additional information that is expected to make the fits more meaningful.

The lsqnonneg function

The present fitting problem belongs to the broader category of *quadratic programming* or *quadratic optimization*, for which numerous methods have been developed in recent years. A much-applied one-pass algorithm is based on the *Kuhn-Tucker theorem* (Lawson & Hanson, 1974), but it is too involved to be reviewed here in full. Let it suffice to mention that it has been implemented in Matlab as the function named the *lsqnonneg*. Below, the variables $\mathbf{k}$, $\mathbf{M}$, and $\mathbf{x}$ must be understood as being defined in the Matlab format. Then we obtain the weight vector $\mathbf{k}$ as

$$\mathbf{k} = \mathrm{lsqnonneg}(\mathbf{M}, \mathbf{x}) \ . \tag{I.3}$$

The *lsqnonneg* function can be computed, and the result will be meaningful, for an *arbitrary rank* of the matrix $\mathbf{M}$. Nonetheless it has to be admitted that there exists a rare theoretical case where the optimal solution is not *unique*. This case occurs, if some of the $\mathbf{m}_i$ *in the final optimal mixture* are *linearly dependent*. In practice, if the input data items to the SOM are stochastic, the probability for the optimal solution being not unique is negligible. At any rate, the locations of the nonzero weights are unique even in this case!

Description of a document by a linear mixture of SOM models

The following analysis applies to most of the SOM applications. Here it is exemplified by textual data bases.

In text analysis, one possible task is to find out whether a text comes from different sources, whereupon its word histogram is expected to be a linear mixture of other known histograms.

The text corpus used in this experiment was taken from a collection published by the Reuters corporation. No original documents were made available; however, Lewis et al. (2004), who have prepared this corpus for benchmarking purposes, have preprocessed the textual data, removing the stop words and reducing the words into their stems. Our work commenced with the ready word histograms. J. Salojärvi from our laboratory selected a 4000-document subset from this preprocessed corpus, restricting only to such articles that were assigned to one of the following classes:

1. Corporate-Industrial.
2. Economics and Economic Indicators.
3. Government and Social.
4. Securities and Commodities Trading and Markets.

There were 1000 documents in each class. Salojärvi then picked up those 1960 words that appeared at least 200 times in the selected texts. In order to carry out *statistically independent experiments*, a few documents were set aside for testing. The 1960-dimensional word histograms were weighted by factors used by Manning and Schütze [3]. Using the weighted word histograms of the rest of the 4000 documents as input, a 2000-node SOM was constructed.

Fig. I.2 shows the four distributions of the hits on the SOM, when the input items from each of the four classes were applied separately to the SOM. It is clearly discernible that the map is *ordered*, i.e., the four classes of documents are segregated to a reasonable accuracy, and the mappings of classes 1, 3, and 4 are even singly connected, in spite of their closely related topics.

Fig. I.3 shows a typical example, where a linear mixture of SOM models was fitted to a new, unknown document. The values of the weighting coefficients k_i in the mixture are shown by dots with relative intensities of color in the due positions of the SOM models. It is to be emphasized that this fitting procedure also defines the optimal *number* of the nonzero coefficients. In the experiments with large document collections, this number was usually very small, less than a per cent of the number of models.

When the models fall in classes that are known a priori, the weight of a model in the linear mixture also indicates the *weight of the class label associated with that model*. Accordingly, by summing up the weights of the various types of class labels one then obtains the *class-affiliation* of the input with the various classes.

References

[1] C.L. Lawson and R.J. Hanson, *Solving Least-Squares Problems*, Englewood Cliffs, NJ: Prentice-Hall, 1974

[2] D.D. Lewis, Y. Yang, T.G. Rose, and T. Li, "RCV1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.* vol. 5, pp.361-397, 2004.

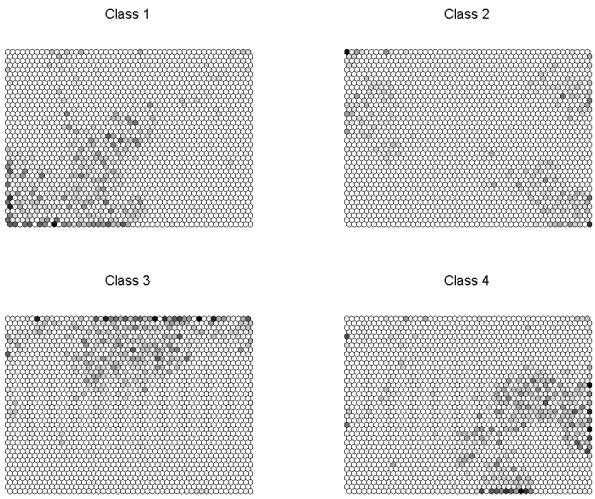

Figure I.2: Mapping of the four Reuters document classes onto the SOM. The densities of the "hits" are shown by shades of gray.

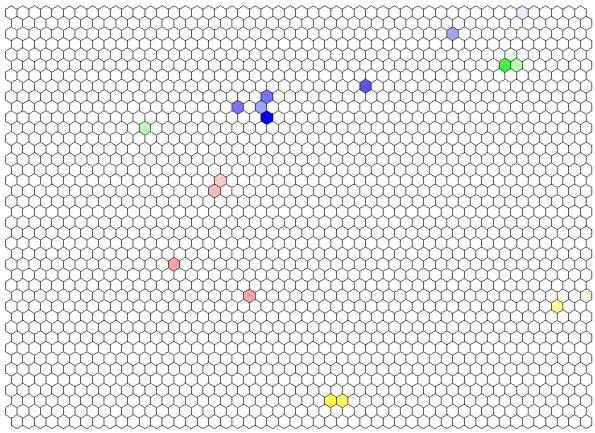

Figure I.3: A linear mixture of SOM models fitted to a new, unknown document. The weighting coefficients k_i in the mixture are shown by using a coloring with a relative saturation of the due models. Red: Class 1. Green: Class 2. Blue: Class 3. Yellow: Class 4.

[3] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press, 1999.

B. Independent variable group analysis

Krista Lagus, Antti Honkela, Jeremias Seppä, Paul Wagner

Independent variable group analysis (IVGA) [1, 2] is a principle for grouping observed input variables so that mutual dependences between variables are strong within a group and weak between groups.

In problems with a large number of diverse observations there are often groups of input variables that have strong mutual dependences within the group but which can be considered practically independent of the input variables in other groups. It can be expected that the larger the problem domain, the more independent groups there are. Estimating a model for each independent group separately produces a more compact representation than applying the model to the whole set of variables. Compact representations are computationally beneficial and, moreover, offer better generalization.

Usually such variable grouping is performed by a domain expert, prior to modeling with automatic, adaptive methods. As expert knowledge may be unavailable, or expensive and time-consuming, automating the task can considerably save resources. The IVGA is a practical, efficient and general approach for obtaining compact representations that can be regarded as sparse codes, as well. Software packages implementing all the presented algorithms are available at `http://www.cis.hut.fi/projects/ivga/`.

The IVGA project is a collaboration with Dr. Esa Alhoniemi (University of Turku) and Dr. Harri Valpola (Helsinki University of Technology, Laboratory of Computational Engineering).

The IVGA algorithm

Any IVGA algorithm consists of two parts, (1) grouping of variables, and (2) construction of an independent model for each variable group. A variable grouping is obtained by comparing models under different groupings using a suitable cost function. In principle any model can be used, if the necessary cost function is derived for the model family.

A practical grouping algorithm for implementing the IVGA principle was first presented in [1]. The method used vector quantizers (VQs) learned with variational Bayesian methods [3] to model the individual groups.

In more recent work [2] we have shown that the variational Bayesian approach is approximately equivalent to minimizing the mutual information or multi-information between the groups. Additionally the modelling algorithm was extended to a finite mixture model that can handle mixed data consisting of both real valued and nominal variables.

Agglomerative grouping algorithm

In addition to the regular combinatorial grouping algorithm corresponding to regular clustering, we have developed an agglomerative IVGA (AIVGA) algorithm for hierarchical grouping of variables [4, 5].

The agglomerative algorithm provides a hierarchical grouping of the variables by starting from singleton groups and merging them iteratively. This both provides a hierarchical view of the variable dependencies as well as a simple greedy deterministic algorithm for solving the ordinary IVGA problem. Experiments reported in [5] show that this method can greatly simplify application of IVGA to practical problems.

Application to computational biology

In [6], IVGA was used to find independent groups of genes or gene regulatory modules from measurements of transcription factor protein binding to different genes in the DNA. The independence of these modules was verified by studying the estimated mutual information of gene expression measurements from mutant organisms with different genes in the discovered modules knocked out. The modules found by IVGA were found to be more meaningful than those discovered by conventional clustering methods.

References

[1] K. Lagus, E. Alhoniemi, and H. Valpola, "Independent variable group analysis," in *Proc. Int. Conf. on Artificial Neural Networks - ICANN 2001*, ser. LNCS, vol. 2130. Vienna, Austria: Springer, 2001, pp. 203–210.

[2] E. Alhoniemi, A. Honkela, K. Lagus, J. Seppä, P. Wagner, and H. Valpola, "Compact modeling of data using independent variable group analysis," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1762–1776, 2007.

[3] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. Jordan, Ed. Cambridge, MA, USA: The MIT Press, 1999, pp. 105–161.

[4] A. Honkela, J. Seppä, and E. Alhoniemi, "Agglomerative independent variable group analysis," in *Proc. 15th European Symposium on Artificial Neural Networks (ESANN 2007)*, Bruges, Belgium, 2007, pp. 55–60.

[5] A. Honkela, J. Seppä, and E. Alhoniemi, "Agglomerative independent variable group analysis," *Neurocomputing*, 2008, doi:10.1016/j.neucom.2007.11.024.

[6] J. Nikkilä, A. Honkela, and S. Kaski, "Exploring the independence of gene regulatory modules," in *Proc. Workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology*, J. Rousu, S. Kaski, and E. Ukkonen, Eds., Tuusula, Finland, 2006, pp. 131–136.

C. Analysis of discrete diffusion scale-spaces

Ramūnas Girdziušas

Taking averages of observations is the most basic method to make inferences in the presence of uncertainty. In late 1980's, this simple idea has been extended to the principle of *successively average less where the change is faster*, and applied to the problem of revealing a signal with jump discontinuities in additive noise.

Successive averaging results in a family of signals with progressively decreasing amount of details, which is called the *scale-space* and further conveniently formalized by viewing it as a solution to a certain diffusion-inspired evolutionary partial differential equation (PDE). Such a model is known as the *diffusion scale-space*.

Example of linear and nonlinear diffusion scale-spaces are shown in Fig. I.4. Diffusion scale-spaces possess two long-standing problems: (i) *model analysis* which aims at establishing stability and guarantees that averaging does not distort important information, and (ii) *model selection*, such as identification of the optimal scale (diffusion stopping time) given an initial noisy signal and an incomplete model.

This thesis studies both problems in the discrete space and time. Such a setting has been strongly advocated by Lindeberg (1991) and Weickert (1996) among others. The focus of the model analysis part is on necessary and sufficient conditions which guarantee that a discrete diffusion possesses the scale-space property in the sense of sign variation diminishing. Connections with the total variation diminishing and the open problem in a multivariate case are discussed too.

Considering the model selection, the thesis unifies two optimal diffusion stopping principles: (i) the time when the Shannon entropy-based Liapunov function of Sporring and Weickert (1999) reaches its steady state, and (ii) the time when the diffusion outcome has the least correlation with the noise estimate, contributed by Mrazek and Navara (2003). Both ideas are shown to be particular cases of the marginal likelihood inference, which is also communicated in [1]. Moreover, the suggested formalism provides first principles behind such criteria, and removes a variety of inconsistencies. It is suggested that the outcome of the diffusion should be interpreted as a certain expectation conditioned on the initial signal of observations instead of being treated as a random sample or probabilities.

(a) (b)

Figure I.4: Example of linear (a) and nonlinear (b) diffusion scales-spaces. They consist of signals which are indexed by the scale value t.

This removes the need to normalize signals, and it also better justifies application of the correlation criterion.

As an example, the following improvement to the existing results can be mentioned. Let the generalized Laplacian $\mathbf{B} \in \mathbb{R}^{n \times n}$ be defined as:

$$\mathbf{B} \equiv - \begin{pmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix} \begin{pmatrix} b_1 & & \\ & \ddots & \\ & & b_{n+1} \end{pmatrix} \begin{pmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}^T , \tag{I.4}$$

where the first matrix is bidiagonal and it contains n rows and $n+1$ columns, whereas the second matrix is diagonal of size $n+1$. Undesignated elements are zeroes.

Theorem 1 *Let $Q_{k,n}$ denote the totality of $\binom{n}{k}$ increasing sequences of integers, each taken from $\{1, 2, \ldots, n\}$ and being of length k. Given any sequence $\boldsymbol{\omega} \in Q_{k,n}$, divide it into r groups of connected indices:*

$$\boldsymbol{\omega} = \{\underbrace{\omega_1, \ldots, \omega_{\nu_1}}_{1st\ group}, \underbrace{\omega_{\nu_1+1}, \ldots, \omega_{\nu_2}}_{2nd\ group}, \ \ldots\ , \underbrace{\omega_{\nu_{r-1}+1}, \ldots, \omega_k}_{r\text{-}th\ group}\} . \tag{I.5}$$

A particular example of $Q_{3,20}$ such as $7, 11, 12$ would produce two groups.

1. *The matrix $(\mathbf{I} - \tau\mathbf{B})^{-1}$ is positive definite if and only if*

$$1 + \sum_{k=1}^{p} \tau^k \sum_{\boldsymbol{\omega} \in Q_{k,p}} \prod_{s=1}^{r} \sum_{i=\nu_{s-1}}^{\nu_s+1} \prod_{\substack{j=\nu_s-1 \\ j \neq i}}^{\nu_s+1} b_j > 0, \quad \textit{for all } p = 1, \ldots, n. \tag{I.6}$$

   In particular, the constraint is satisfied if $b_i > 0$ for all $i = 2, \ldots, n$.

2. *The matrix $\mathbf{I} + \tau\mathbf{B}$ is positive definite if and only if*

$$1 + \sum_{k=1}^{p} \tau^k \sum_{\boldsymbol{\omega} \in Q_{k,p}} \prod_{s=1}^{r} \sum_{i=\nu_{s-1}}^{\nu_s+1} \prod_{\substack{j=\nu_s-1 \\ j \neq i}}^{\nu_s+1} (-b_j) > 0, \quad \textit{for all } p = 1, \ldots, n. \tag{I.7}$$

   In particular, the constraint is satisfied if $0 \leq b_2 \leq \tau^{-1}$, $0 \leq b_{i-1} + b_i \leq \tau^{-1}$ for all $i = 3, \ldots, n$, and $0 \leq b_n \leq \tau^{-1}$.

The proof with a geometric description can be found in [3]. All known previous characterizations either include determinantal quantities, or provide only sufficient conditions such as the case of positive diagonal elements with a diagonal dominance, which follows from Gershgorin's circles. Theorem 1 has been related to the scale-space analysis in [2].

References

[1] R. Girdziušas and J. Laaksonen. How marginal likelihood inference unifies entropy, correlation and snr-based stopping in nonlinear diffusion scale-spaces. In I. S. Kweon Y. Yagi, S. B. Kang and H. Zha, editors, *Proc. of 8th Asian Conf. on Computer Vision*, volume 4843 of *Lecture Notes in Computer Science*, pages 811–820, 2007.

[2] R. Girdziušas. *Stability and Inference in Discrete Diffusion Scale-Spaces.* Doctoral thesis, Helsinki University of Technology, 2008.

[3] R. Girdziušas and J. Laaksonen. When is a discrete diffusion a scale-space? In A. Shashua D. Metaxas, B. C. Vemuri and H. Shum, editors, *Proc. of 11th IEEE Int. Conf. on Computer Vision*, page 6. IEEE, 2007.

D. Feature selection for steganalysis

Yoan Miche, Amaury Lendasse, Patrick Bas and Olli Simula

Steganography has been known and used for a very long time, as a way to exchange information in an unnoticeable manner between parties, by embedding it in another, apparently innocuous, document. For example, during the 80's, Margaret Thatcher decided to have each word processor of the government's administration members changed with an unique word spacing for each, giving a sort of "invisible signature" to documents. This was done to prevent the continuation of sensitive government information leaks.

Nowadays steganographic techniques are mostly used on digital contents. The online newspaper, Wired News, reported in one of its articles on steganography that several steganographic contents have been found on web-sites with very large image database such as eBay.

Most of the time research about steganography is not as much to hide information, but more to detect that there is hidden information. This "reverse" part of the steganography is called steganalysis and is specifically aimed at making the difference between genuine documents, and steganographied – called stego – ones. Consequently, steganalysis can be seen as a classification problem where the goal is to build a classifier able to distinguish these two sorts of documents.

During the steganographic process, a message is embedded in an image so that it is as undetectable as possible. Basically, it uses several heuristics in order to guarantee that the statistics of the stego content (the modified image) are as close as possible to the statistics of the original one. Afterwards, steganalysis techniques classically use features extracted from the analyzed image and an appropriately trained classifier to decide whether the image is genuine or not.

In our work, a widely used and known set of 193 image features has been used. Theses features consider statistics of JPEG compressed images such as histograms of DCT coefficients for different frequencies, histograms of DCT coefficients for different values, global histograms, blockiness measures and co-occurrence measures. The main purpose of this high number of features is to obtain a model able to detect about any steganographic process.

The usual process in steganalysis is then to train a classifier according to the extracted features. Consequently a set of 193 features for each image of the database is obtained, giving an especially high dimensionality space for classifiers to work on. Earlier research about these high dimensionality spaces has shown that a lot of issues come out when the number of features is as high as this one.

The main idea behind the carried out work [1, 2, 3, 4] is to give insights on proper handling and use of such high dimensionality datasets; indeed, these are very common in the steganography/steganalysis field and users tend not to respect basic principles (for example having a sufficient number of samples regarding the dimensionality of the problem). In the framework of an international thesis co-agreement between the GIPSA-lab in Institut National Polytechnique de Grenoble (France) and ICS laboratory in Helsinki University of technology, Yoan Miche (GIPSA-lab, ICS) along with Patrick Bas (GIPSA-lab) and Amaury Lendasse (ICS), his advisors, as well as Olli Simula (ICS), his supervisor, developed a methodology for handling these datasets; this methodology is used to determine a sufficient number of images for effective training of a classifier in the obtained high-dimensional space, and use feature selection to select most relevant features for the

desired classification. Dimensionality reduction managed to reduce the original 193 features set by a factor of 13, with overall same performance.

By the use of a Monte-Carlo technique on up to 4000 images, it has been shown that such numbers of images are sufficient for stable results when having a set of 193 features extracted from all images. In the experiments, dimensionality reduction managed to reduce the number of required features to 14, while keeping roughly the same classification results. Computational time is thus greatly improved, divided by about 11. Also, further analysis becomes again possible with this low number of features: conclusions and precisions about the steganographic scheme can be inferred from the obtained feature set.

References

[1] Y. Miche and P. Bas and A. Lendasse and O. Simula and C. Jutten, *Avantages de la Sélection de Caractéristiques pour la Stéganalyse*, in GRETSI 2007, Groupe de Recherche et d'Etudes du Traitement du Signal et des Images, Troyes, France, September 11-13 2007.

[2] Y. Miche and P. Bas and A. Lendasse and C. Jutten and O. Simula, *Advantages of Using Feature Selection Techniques on Steganalysis Schemes*, in IWANN'07: International Work-Conference on Artificial Neural Networks, San Sebastian, Spain, June 20-22 2007.

[3] Y. Miche and P. Bas and A. Lendasse and C. Jutten and O. Simula, *Extracting Relevant Features of Steganographic Schemes by Feature Selection Techniques*, in Wacha'07: Third Wavilla Challenge, Saint Malo, France, June 14 2007.

[4] Y. Miche and B. Roue and P. Bas and A. Lendasse, *A Feature Selection Methodology for Steganalysis*, in MRCS06, International Workshop on Multimedia Content Representation, Classification and Security, Istanbul, Turkey, September 11-13 2006.

E. Adaptive committee techniques

Matti Aksela, Jorma Laaksonen, Erkki Oja

Combining the results of several classifiers can improve performance because in the outputs of the individual classifiers the errors are not necessarily overlapping. Also the combination method can be adaptive. The two most important features of the member classifiers that affect the committee's performance are their individual error rates and the diversity of the errors. The more different the mistakes made by the classifiers, the more beneficial the combination of the classifiers can be.

Selecting member classifiers is not necessarily simple. Several methods for classifier diversity have been presented to solve this problem. In [2] a scheme weighting similar errors made in an exponential fashion, the Exponential Error Count method, was found to provide good results. Still, the best selection of member classifiers is highly dependent on the combination method used.

We have experimented with several adaptive committee structures. Two effective methods have been the Dynamically Expanding Context (DEC) and Class-Confidence Critic Combining (CCCC) schemes. The DEC algorithm was originally developed for speech recognition purposes. The main idea is to determine just a sufficient amount of context for each individual segment so that all conflicts in classification results can be resolved. In the DEC committee, the classifiers are initialized and ranked in the order of decreasing performance. Results of the member classifiers are used as a one-sided context for the creation of the DEC rules. Each time a character is input to the system, the existing rules are searched through. If no applicable rule is found, the default decision is applied. If the recognition was incorrect, a new rule is created.

In our CCCC approach the main idea is to try to produce as good as possible an estimate on the classifier's correctness based on its prior behavior for the same character class. This is accomplished by the use of critics that assign a confidence value to each classification. The confidence value is obtained through constructing and updating distribution models of distance values from the classifier for each class in every critic. These distribution models are then used to extract the needed confidence value, based on prior results in addition to the sample being processed. The committee then uses a decision mechanism to produce the final output from the input label information and critic confidence values. In our earlier experiments the adaptive committee structures have been shown to be able to improve significantly on their members' results.

Also classifiers that are adaptive in themselves can be combined using an adaptive committee. Experiments have shown that while making a single classifier adaptive does produce on average the best gains when used alone, the addition of another layer of adaptation, when implemented in a robust fashion, can produce even better results than either method alone [1].

References

[1] Matti Aksela and Jorma Laaksonen. Adaptive combination of adaptive classifiers for on-line handwritten character recognition. *Pattern Recognition Letters*, 28(1):136–143, 2007.

[2] Matti Aksela and Jorma Laaksonen. Using diversity of errors for selecting members of a committee classifier. *Pattern Recognition*, 39(4):608–623, 2006.