ASSESSMENT OF THE EFFECT OF NOISE ON AN UNSUPERVISED FEATURE SELECTION METHOD FOR GENERATIVE TOPOGRAPHIC MAPPING

Alfredo Vellido and Jorge S. Velazco
Department of Computing Languages and Systems (LSI), Technical University of Catalonia (UPC), Barcelona, Spain

Keywords: Unsupervised Feature Selection, Feature Relevance Determination, Generative Topographic Mapping, clustering, uninformative noise.

Abstract: Unsupervised feature relevance determination and feature selection for dimensionality reduction are important issues in many clustering problems. An unsupervised feature selection method for general Finite Mixture Models was recently proposed and subsequently extended to Generative Topographic Mapping (GTM), a nonlinear manifold learning constrained mixture model for data clustering and visualization. Some of the results of a previous preliminary assessment of this method for GTM suggested that its performance may be affected by the presence of uninformative noise in the dataset. In this brief study, we test in some detail such limitation of the method.

1 INTRODUCTION

Statistical Machine Learning (SML) provides a unified principled framework for machine learning methods and helps to overcome some of their limitations. Embedding probability theory into machine learning techniques has important modeling implications. For instance, it requires modeling assumptions, including the specification of prior distributions, to be made explicit; it also automatically satisfies the likelihood principle and provides a natural framework to handle uncertainty.

An example of SML can be found in Finite Mixture Models (FMM), which are flexible and robust methods for multivariate data clustering (McLachlan and Peel, 1998). The addition of visualization capabilities would benefit these models in many application scenarios, helping to provide intuitive cues about data structural patterns. One way to endow FMM with data visualization is by constraining the mixture components to be centered in a low-dimensional manifold embedded into the multivariate data space, as in Generative Topographic Mapping (GTM) (Bishop et al., 1999). This is a non-linear, neural network-inspired manifold learning model for simultaneous data clustering and visualization.

The interpretability of the clustering results provided by GTM becomes difficult when the analyzed data sets consist of a large number of features. This limitation can be overcome with methods to estimate the ranking of the data features according to their relative relevance, leading to feature selection (FS). The research on unsupervised FS is scarce in comparison to that for supervised models, despite the fact that FS becomes a paramount issue in many clustering problems. A description of the problem in terms of a reduced subset of relevant features would improve the interpretability of the clusters obtained by unsupervised methods.

An important advance on unsupervised FS for Finite Mixture Models was presented in (Law et al., 2004) and recently extended to GTM (the FRD-GTM model) in (Vellido et al., 2006) and to one of its variants for time series analysis (FRD-GTM-TT) in (Olier and Vellido, 2006). This method was preliminarily assessed in (Vellido, 2006), where some of the results suggested that the performance of the method may be degraded by the presence of uninformative noise, which would obscure the underlying cluster structure of the data and, therefore, mislead an unsupervised feature relevance estimation method. In this brief study, we provide evidence of the limitations of the method through controlled experiments using synthetic data.

The remaining of the paper is organized as follows. First, brief introductions to the standard Gaussian GTM and its extension for Feature Relevance Determination (FRD) are provided in section 2. This is

Vellido A. and S. Velazco J. (2008).
ASSESSMENT OF THE EFFECT OF NOISE ON AN UNSUPERVISED FEATURE SELECTION METHOD FOR GENERATIVE TOPOGRAPHIC MAPPING.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - AIDSS*, pages 424-430
DOI: 10.5220/0001681204240430

followed, in section 3, by a description of the experimental settings and, in section 4, by a presentation and discussion of the results. The paper closes with a brief summary of conclusions.

2 FEATURE RELEVANCE DETERMINATION FOR GTM

2.1 The Standard GTM Model

The neural network-inspired GTM is a manifold learning model with sound foundations in probability theory. It performs simultaneous clustering and visualization of the observed data through a nonlinear and topology-preserving mapping from a visualization latent space in $\Re^L$ (with L being usually 1 or 2 for visualization purposes) onto a manifold embedded in the $\Re^D$ space, where the observed data reside. For each feature d, the functional form of this mapping is the generalized linear regression model $y_d(\mathbf{u},\mathbf{W}) = \sum_m^M \phi_m(\mathbf{u})\, w_{md}$, where ϕ_m is one of M basis functions, defined here as spherically symmetric Gaussians, generating the non-linear mapping from a latent vector $\mathbf{u}$ to the manifold in $\Re^D$. The matrix $\mathbf{W}$ of adaptive weights w_{md} explicitely defines this mapping.

The prior distribution of $\mathbf{u}$ in latent space is constrained to form a uniform discrete grid of K centres. A density model in data space is therefore generated for each component k of the mixture, which, assuming that the observed data set $\mathbf{X}$ is constituted by N independent, identically distributed (i.i.d.) data points $\mathbf{x}_n$, leads to the definition of a complete log-likelihood in the form:

$$L(\mathbf{W},\beta|\mathbf{X})= \sum_{n=1}^N \ln\left\{ \frac{1}{K}\sum_{k=1}^K \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\beta/2\|\mathbf{y}_k-\mathbf{x}_n\|^2\right\}\right\} \quad (1)$$

where $\mathbf{y}_k$ is a reference or prototype vector consisting of elements ($y_{dk} = \sum_m^M \phi_m(\mathbf{u}_k)\, w_{md}$), which are an instantiation of the generalized linear regression model described above. From Eq. (1), the adaptive parameters of the model, which are $\mathbf{W}$ and the common inverse variance of the Gaussian components, β, can be optimized by maximum likelihood (ML) using the Expectation-Maximization (EM) algorithm. Details can be found in (Bishop et al., 1999).

2.2 The FRD-GTM

In this paper, unsupervised feature relevance is understood as the likelihood of a feature being responsible for generating the data cluster structure. Therefore, relevant features will be those which better separate the natural clusters in which the data are structured. Moreover, we are interested in unsupervised feature selection methods that are suitable for clustering models that also provide data visualization. With that in mind, the FRD technique was defined for the GTM model in (Vellido et al., 2006). For the unsupervised GTM clustering model, relevance is defined through the concept of saliency.

The FRD problem was investigated for GTM in (Vellido et al., 2006). Feature relevance in this unsupervised setting is understood as the likelihood of a feature being responsible for generating the data cluster structure and it is quantified through the concept of saliency. Formally, the saliency of feature d can be defined as $\rho_d = P(\eta_d = 1)$, where $\eta=(\eta_1,\ldots,\eta_D)$ is a set of binary indicators that can be integrated in the EM algorithm as missing variables. A value of $\eta_d = 1$ $(\rho_d = 1)$ indicates that feature d has the maximum possible relevance. According to this definition, the FRD-GTM mixture density can be written as:

$$p(\mathbf{x}|\mathbf{W},\beta,\mathbf{w}_0,\beta_0,\rho)= \sum_{k=1}^K \frac{1}{K}\prod_{d=1}^D\left\{\rho_d\, p(x_d|\mathbf{u}_k;\mathbf{w}_d,\beta)+(1-\rho_d)\,q\left(x_d|\mathbf{u}_0;w_{0,d},\beta_{0,d}\right)\right\} \quad (2)$$

where $\mathbf{w}_d$ is the vector of $\mathbf{W}$ corresponding to feature d and $\rho \equiv \{\rho_1,\ldots,\rho_D\}$. A feature d will be considered irrelevant, with *irrelevance* $(1-\rho_d)$, if $p(x_d|\mathbf{u}_k;\mathbf{w}_d,\beta) = q(x_d|\mathbf{u}_0;w_{0,d},\beta_{0,d})$ for all the mixture components k, where q is a common density followed by feature d. Notice that this is like saying that the distribution for feature d does not follow the cluster structure defined by the model. This common component requires the definition of two extra adaptive parameters: $\mathbf{w}_0 \equiv \{w_0,1,\ldots,w_{0,D}\}$ and $\beta_0 \equiv \{\beta_{0,1},\ldots,\beta_{0,D}\}$ (so that $\mathbf{y}_0 = \phi_0(\mathbf{u}_0)\,\mathbf{w}_0$). For fully relevant ($\rho_d \to 1$) features, the common component variance vanishes:$(\beta_{0,d})^{-1} \to 0$. The parameters of the model can, once again, be optimized by ML using the EM algorithm. Detailed calculations can be found in (Vellido, 2005).

3 EXPERIMENTAL SETTINGS

The results of statistically principled models for probability density estimation, such as GTM and its variants, are bound to be affected, in one way or another, by the presence of uninformative noise in the data. Here, we assess such effects on the FRD-GTM model described in the previous section. For that, data with very specific characteristics are required. We use synthetic sets similar to those in (Law et al., 2004) for comparative purposes.

The first synthetic set (hereafter referred to as *synth1*) is a variation on the *Trunk* data set used in (Law et al., 2004)), and was designed for its 10 features to be in decreasing order of relevance. It consists of data sampled from two Gaussians $N(\mu_1, \mathbf{I})$ and $N(\mu_2, \mathbf{I})$, where $\left(\mu_1 = 1, \frac{1}{\sqrt{3}}, \ldots, \frac{1}{\sqrt{2d-1}}, \ldots, \frac{1}{\sqrt{19}}\right)$ and $\mu_1 = -\mu_2$. We hypothesize (*H1*) that the feature relevance ranking estimated by FRD-GTM for these data will deteriorate gradually as noise is added to the 10 original features and in proportion to its level. In order to test *H1*, four increasing levels of Gaussian noise, of standard deviations 0.1, 0.2, 0.5, and 1, were added to the 10 original features of *synth1*, for a given sample size. It is also hypothesized (*H2*) that the feature relevance ranking will deteriorate as we add new noisy features and in proportion to their level of noise. In order to test *H2*, 5 and 10 dummy features consisting of Gaussian noise of standard deviations 0.1, 0.2, 0.5, and 1, were, in turn, added to the 10 original features.

The second dataset (hereafter referred to as *synth2*) consists of two features defining four neatly separated Gaussian clusters with centres located at $(0,3), (1,9), (6,4)$ and $(7,10)$; they are meant to be relatively relevant in contrast to any added noise. In a first experiment, noise of different levels was added to the first two features, while 4 extra noise features were added to those two. Several other experiments, similar to the ones devised for *synth1* were designed to further test *H2*.

The FRD-GTM parameters $\mathbf{W}$ and $\mathbf{w}_0$ were initialized with small random values sampled from a normal distribution. Saliencies were initialized at $\rho_d = 0.5, \forall d, d = 1, \ldots, D$. The grid of GTM latent centres was fixed to a square layout of 3×3 nodes (i.e., 9 constrained mixture components). The corresponding grid of basis functions ϕ_m was fixed to a 2×2 layout.

4 EXPERIMENTAL RESULTS AND DISCUSSION

The experiments outlined in the previous section aim to assess the effect of the presence of uninformative noise on the performance of FRD-GTM in the process of unsupervised feature relevance estimation.

In the experiments reported in Figure 1, four levels of Gaussian noise of increasing level were added to a sample of 1,000 points of *synth1*. The FRD-GTM is shown to behave robustly even in the presence of a substantial amount of noise, although its performance deteriorates significantly for noise of standard deviation = 1, as reflected in the breach of the expected monotonic decrease of the mean feature saliencies. It is also true that, comparing these results with those in Figure 2 (in which no noise was added to *synth1*), the most relevant feature is not so close to a saliency of 1. *H1* is, therefore, partially supported by these results.

The FRD ranking results for the second experiment, using the 10 original features of *synth1* plus 5 Gaussian noise features, are shown in Figure 2. For all levels of noise, the relevance (in the form of estimated saliency) of the original features $(1 \rightarrow 10)$ is reasonably well estimated: the saliency for the first feature is close to 1 with almost full certainty (very small vertical bars) and, overall, the expected monotonic decrease of the mean feature saliencies is preserved, although breaches of such monotonicity can also be observed. The saliencies estimated for the 5 added Gaussian noise features are regularly estimated to be small. Interestingly, the increase in the level of noise does not seem to affect the performance of the FRD method in any significant way: the differences between the saliencies of the 10 original variables and the 5 noisy ones stay roughly the same and the decreasing relevance for the 10 original variables does not vary substantially. According to these results, *H2* is not supported at this stage.

The FRD ranking results for the third experiment, using the 10 original features of *synth1* plus 10 Gaussian noise features are shown in Figure 3. Once again, and for all levels of noise, the relevance of the 10 original features shows, overall, the expected monotonic decrease of the mean feature saliencies, with some breaches of monotonicity. This time, the saliencies estimated for the 10 added Gaussian noise features are not that clearly small in comparison to those estimated for the 10 original ones. In summary, the decreasing relevance for the 10 original variables does not vary substantially, and the differences between the saliencies of the 10 original features and the 5 noisy ones stay roughly the same regardless the noise level. Nevertheless, the FRD method seems to be affected by the increase in number of the noisy features. According to these results, *H2* is only partially supported.

The FRD-GTM is shown to behave with reasonable robustness when noise is added to the first two features of *synth2*, as shown in Figure 4. As in the case of *synth1*, its performance deteriorates significantly for high levels of noise. Comparing these results with those in Figures 5 and 6 (in which no noise was added to the first two features), the overall deterioration becomes evident. *H1* is again partially supported by these results.

The FRD ranking results for the experiments using the 2 original features of *synth2* plus either 7 or 10 Gaussian noise features are shown, in turn, in Figures 5 and 6. This is clearly a far easier problem for the FRD method. Regardless the level of noise and the number of added noisy features, FRD-GTM consistently estimates the first 2 features to be the most relevant. Furthermore, the differences between the saliencies estimated for the first 2 features and the added (7 or 10) noisy ones stay roughly the same. In contrast with the results obtained in the experiments with *synth1*, the estimated saliencies for all noisy features are low and quite similar. Our research hypothesis *H2* is not supported by these results.

5 CONCLUSIONS

In this paper, the effects of the presence of noise on a method of unsupervised feature relevance determination for the manifold learning GTM model, have been investigated in some detail.

The FRD-GTM has been shown to behave with reasonable robustness even in the presence of a fair amount of noise. It was first hypothesized that the feature relevance ranking would deteriorate as we add noise to the existing features and in proportion to the level of that noise. This hypothesis has found only limited experimental support. It was also hypothesized that the feature relevance ranking would deteriorate as we add extra noisy features to the existing ones and in proportion to their number and the level of noise. This second hypothesis has found little experimental support: There is only some evidence that the performance of the FRD method deteriorates as we increase the number of purely noisy features and only if the dataset is complex enough.

This relative weakness of the method in the presence of noise makes it convenient to consider possible strategies for model regularization and, therefore, future research will be devoted the design of methods for automatic and proactive model regularization to prevent or at least limit the negative effect of data overfitting on the FRD method for GTM. Some of such methods have already been designed for the standard GTM formulation (Bishop et al., 1998; Vellido et al., 2003) and could be extended to FRD-GTM. Alternatively, regularization could be accomplished through a reformulation of the GTM within a variational Bayesian theoretical framework (Olier and Vellido, 2008). Again, this could be extended to accomodate FRD.

Future research should extend the experimental design to include a wider variety of artificial data sets of different characteristics. It should also address the design of strategies for adaptive model regularization for FRD-GTM. Such kind of strategy would automatically regulate the level of map smoothing necessary to avoid the model fitting the noise in the data, i.e. data overfitting.

ACKNOWLEDGEMENTS

Alfredo Vellido is a researcher within the Ramón y Cajal program of the Spanish Ministry of Education and Science (MEC) and acknowledges funding from the MEC I+D project TIN2006-08114.

REFERENCES

Bishop, C., Svensén, M., and Williams, C. (1998). Developments of the generative topographic mapping. In *Neurocomputing*. 21(1-3), pp. 203-224.

Bishop, C., Svensén, M., and Williams, C. (1999). Gtm: The generative topographic mapping. In *Neural Computation*. 10(1), pp. 215–234.

Law, M., Figueiredo, M., and Jain, A. (2004). Simultaneous feature selection and clustering using mixture models. In *IEEE T. Pattern Anal*. 26(9), pp. 1154–1166.

McLachlan, G. and Peel, D. (1998). *Finite mixture models*. John Wiley-Sons, New York.

Olier, I. and Vellido, A. (2006). Time series relevance determination through a topology-constrained hidden markov model. In *Proc. of the 7th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2006)*. LNCS 4224, 40-47. Burgos, Spain.

Olier, I. and Vellido, A. (2008). On the benefits for model regularization of a variational formulation of gtm. In *in Proceedings of the International Joint Conference on Neural Networks (IJCNN 2008)*. in press.

Vellido, A. (2005). Preliminary theoretical results on a feature relevance determination method for generative topographic mapping. In *Technical Report LSI-05-13-R*. Universitat Politecnica de Catalunya, Barcelona, Spain.

Vellido, A. (2006). Assessment of an unsupervised feature selection method for generative topographic mapping. In *16th International Conference on Artificial Neural Networks*. LNCS 4132, 361-370. Athens, Greece.

Vellido, A., El-Deredy, W., and Lisboa, P. (2003). Selective smoothing of the generative topographic mapping. In *IEEE T. Neural Network*. 14(4), pp. 847-852.

Vellido, A., Lisboa, P., and Vicente, D. (2006). Robust analysis of mrs brain tumour data using t-gtm. In *Neurocomputing*. 69(7–9), pp. 754–768, 2006.

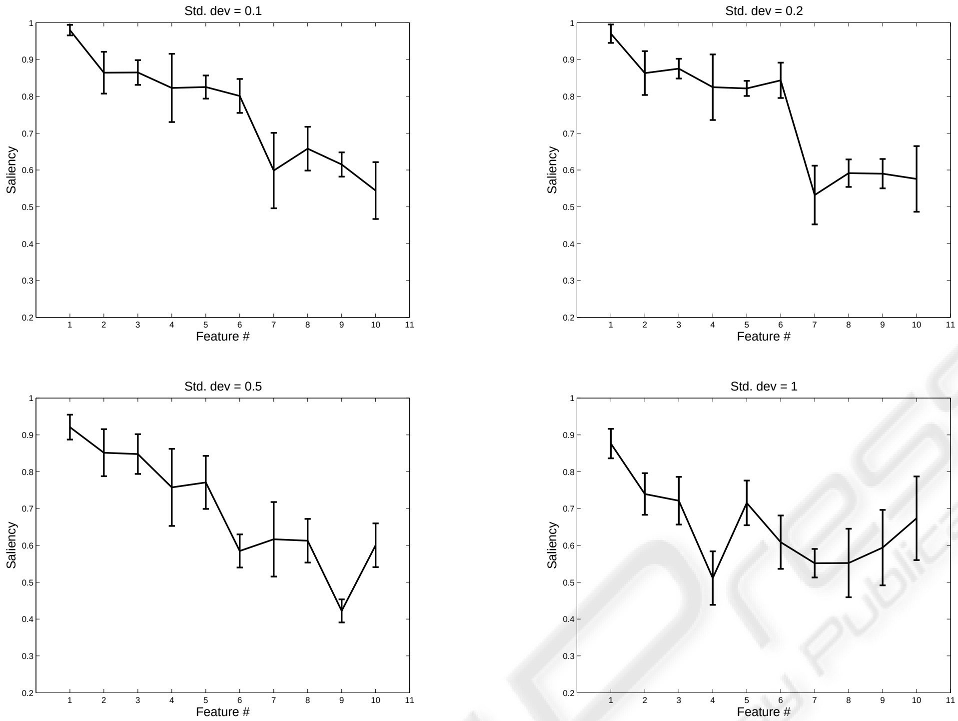

Figure 1: Experiments with a sample of 1,000 points from *synth1*, to which different levels of Gaussian noise (indicated in the plot titles) were added to the existing features. Mean saliencies ρ_d for the 10 features. The bars span from the mean minus to the mean plus one standard deviation of the saliencies over 20 runs of the algorithm.

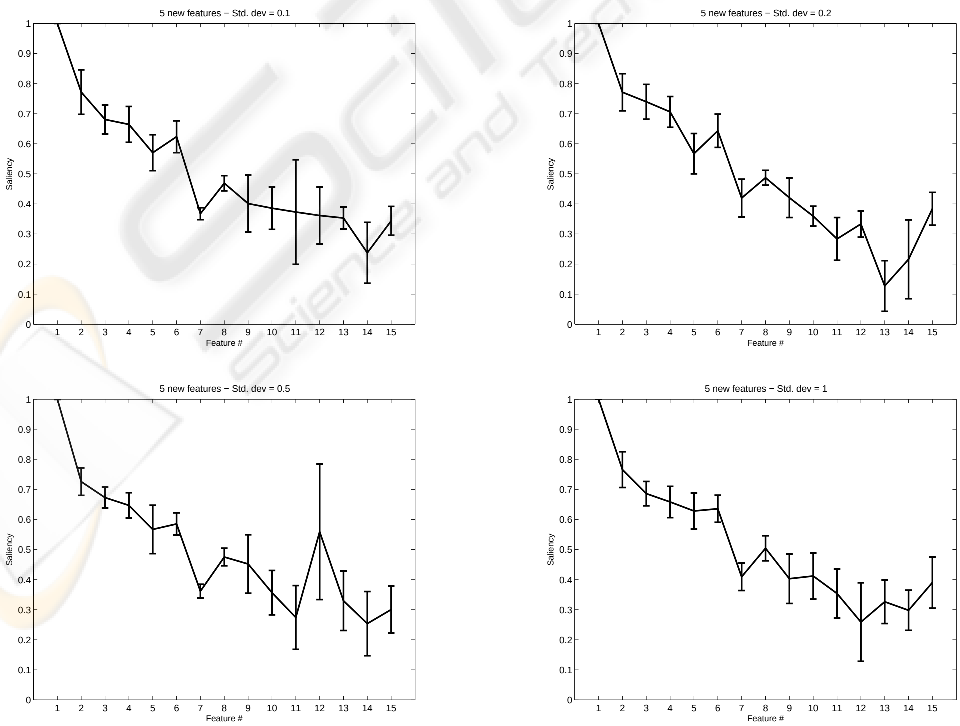

Figure 2: Experiments with a sample of 1,000 points from *synth1*, to which 5 extra noise features ($11 \rightarrow 15$) of different noise levels (indicated in the plot titles) were added. Representation as in Figure 1.

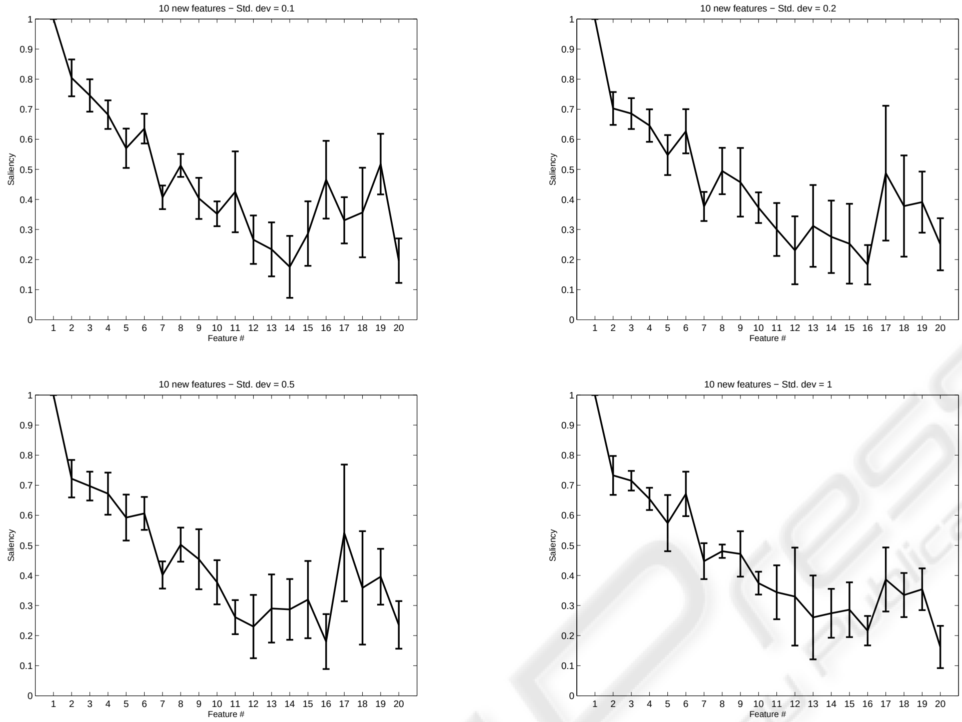

Figure 3: Experiments with a sample of 1,000 points from *synth1*, to which 10 extra noise features (11 → 20) of different noise levels (indicated in the plot titles) were added. Representation as in Figure 1.

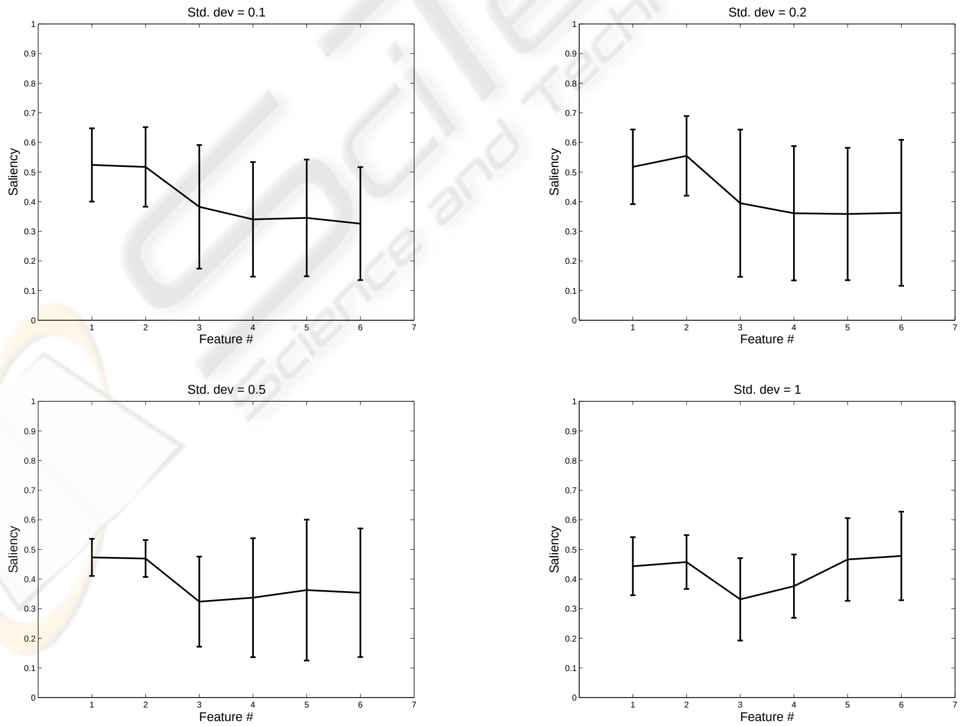

Figure 4: Experiments with a sample of 1,000 points from *synth2*, to which noise of different levels (indicated in the plot titles) were added. Four extra noise features (3 → 6) of the same noise levels were added. Representation as in previous figures.

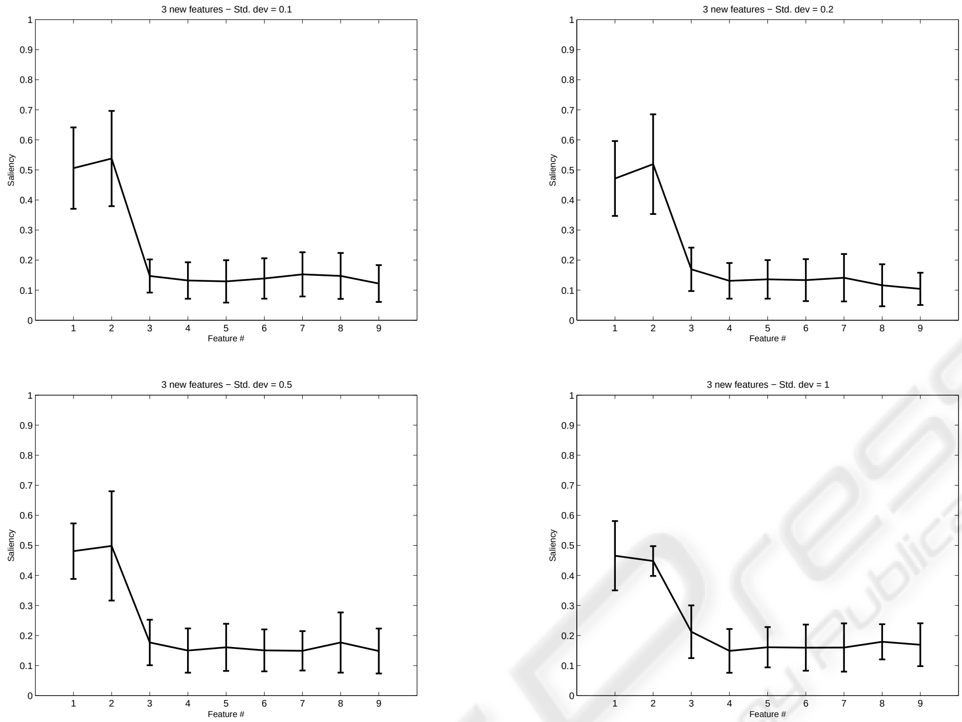

Figure 5: Experiments with a sample of 1,000 points from *synth2*, to which 7 extra noise features ($3 \rightarrow 9$) of different noise levels (indicated in the plot titles) were added. Representation as in previous figures.

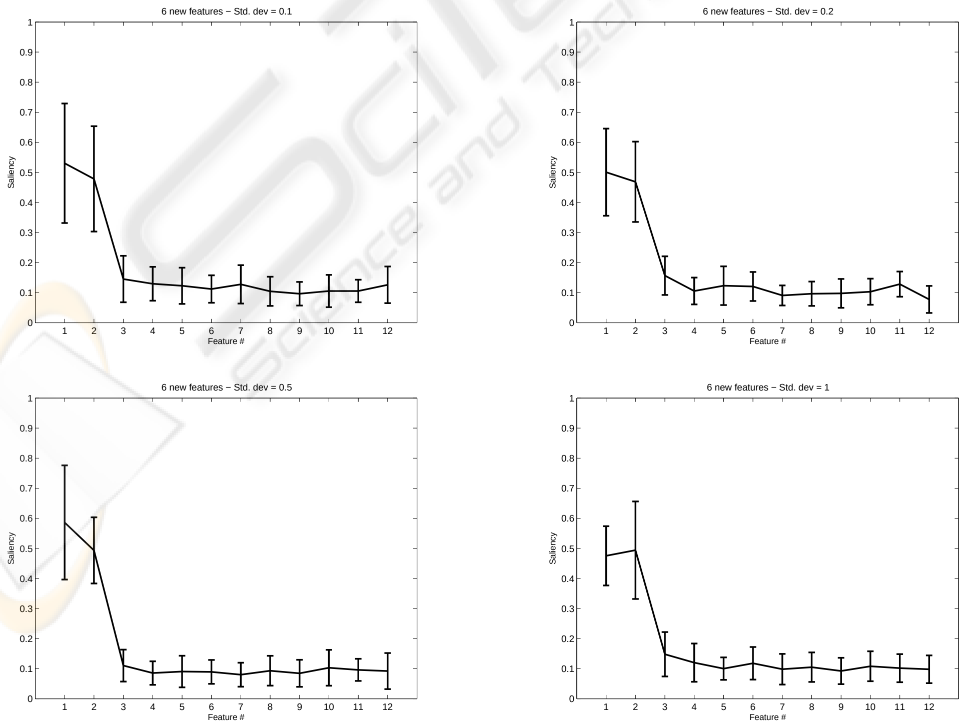

Figure 6: Experiments with a sample of 1,000 points from *synth2*, to which 10 extra noise features ($3 \rightarrow 12$) of different noise levels (indicated in the plot titles) were added. Representation as in previous figures.