18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

A Proposal for Improving Spoken Dialog Systems using Context Information Fusion

I. Chairi, D. Griol, J. García, J. M. Molina
Computer Science Department
University Carlos III de Madrid
Spain
{ikram.chairi, david.griol, jesus.garcia, josemanuel.molina}@uc3m.es

Abstract - ***Context information fusion allow developing systems that provides large quantities of information obtained from network sensors with heterogeneous characteristics, that needs to be efficiently processed. During the last years, this field has become increasingly relevant due to the growing importance and rapid advances in sensor technology that provide information from the environment. Many works has incorporate context into a fusion data framework, showing the improvement of accuracy by integrating context, others proved the usefulness of context to improve classification task by using multiple classifier system. This paper proposes an improvement of a spoken dialog system using context information fusion. An effective application of this approach is described for a spoken dialog system providing railway information.***

Keywords: Context Information Fusion, Multiple classification, Spoken Dialog Systems.

1 Introduction

Information fusion is the process of combing information from a number of different sources to achieve improved accuracies with respect to those achieved by a single source. Multi-source information fusion is widely applied in different field including military application, commercial systems and medical applications. In principle, fusion of data from multisensor provides significant advantages over single source data. The use of multiple types of sensors may increase the accuracy with which a quantity can be observed and characterized [1].

One of the key issues in developing a multisensor data fusion system is what architecture (or framework) should be used. Frameworks however are very difficult to make because of the necessity to consider the domain of interest in their design, and also should be done in a way that can accommodate different applications in the domain-space. Instead, the choice of architecture must balance computing resources, available communication bandwidth, desired accuracy, the capabilities of the sensors, and available funding.

Generally, observational data may be combined at two levels. The first one ("early fusion") is from the raw data (or observation) level to a state vector level where features are extracted from multiple sensor observations, and combined into a single concatenated feature vector which is input to pattern recognition approaches based on neural networks, clustering algorithms, or template methods. The second level of combination ("late fusion") is at the decision level which involves fusion of sensor information, after each sensor has made a preliminary determination of an entity's location, attributes, and identity [1].

Many works on information fusion framework has emerged since the Joint Directors of Laboratories' Data Fusion Sub-Panel (now the Data Fusion Group) introduced a model of data fusion (JDL model) and defined a lexicon with some terms of reference for data fusion. A summary from the literature that address IF framework concepts and issues is presented in [2]. [3], [4] also present a survey of different architectural model used for information fusion which are generally grouped in Information-based model, activity-based model, and role-based model. Of the many possible ways of differentiating among types of data fusion processes, that of JDL has gained the greatest popularity [5].

With the continuing expansion of the domain of application and the increasing complexity of the collected information, intelligent techniques for fusion processing have become a crucial component in information fusion applications. Intelligent systems can improve information fusion based on the context management in order to support decision making.

"Context" is a broad term used in the Information Fusion community, there has been active research on how to represent and exploit context in fusion processes in different ways during the last twenty years. Recent analysis can be found in Special Issues [6], [7] and a survey in [8]. In this field, context can be defined as information that "surrounds" the situation of interest in the estimation process, so it aids in its understanding and decision support, if necessary [9]. Sometimes context is defined in the sense of constraints, as in [10]: "the structured set of variables, external constraints to some (natural or artificial) cognitive process that influences the behavior of that process in the agent(s) under consideration". This approach is in line with approaches to context from the AI community, one of the first approximations to context is due to McCarthy [11], extending logic relations to explicitly include context.

Typical areas in Information Fusion are threat assessment and crisis management, requiring decision-making support about situation. So, Rogova discusses in [12] how context plays a central role in these functions, and [13] propose a formal structure of domain (entities, attributes, relations, etc) with ontologies for reasoning about situations, intent and threats. Context has been also used to detect anomalous situations also, as presented in [14] for a harbor surveillance scenario, including rule-based reasoning to extend tracking data and classify objects according to pre-defined categories defined in ontologies.

Following Steinberg et al [15], the use of context for predicting and understanding situations can be oriented to establishing expectations about the states of individuals, events or situations of interest in decision-making. Besides, relevant contextual variables might not be known a priori so a form of dynamic context discovery should also be carried out as part of the optimization process. This is discussed in [16], with inference methods to select context variables on the basis of their utility in refining explicit problem variables, selecting and fusing contextual information as part of a goal-driven decision process.

Our approach is the exploitation of context information in spoken dialog system providing railway information system. A spoken dialog system try to speech similar to that between humans where dialog manager decides the next action of the system using a classification procedure that could be improved by using the context knowledge.

2 Improvement of classification using context information

2.1 Incorporation of context in classification system

Recently, information fusion has been widely used in the field of classification. In fact, the use of data fusion systems to carry out classification tasks becomes a necessity since the process of data acquisition uses many sources of information. Many works have proved the efficiency of using information fusion for classification. For instance, in [17], the use of multi-sensor data fusion for urban area classification allows to increase the number of classes and to boost the accuracy of the classification. Another example of integration of multisensor information fusion in classification system is presented in [18], where a multi-source data fusion classifier is developed and applied to land cover classification. In these purpose a new classifier able to update framework using contextual information, has shown good promising results.

The idea of using context information for learning a classification system is relatively new. Even if few works deal with this concept yet, it has shown the effectiveness of using context information for classification. For instance, in [19], classification of images in the medical domain is improved by incorporating context. The proposal is considering only the measurements related to an object and its relevant context. To do that, context has to be extracted (manually or with an automatic previous process) so that some form of relevance function is applied to select the relevant data based on context. In [20] context is represented by a network of situations and the work proposes a framework for generic situation acquisition algorithm with an application to video surveillance. Learning of complex domain knowledge is discussed in [21] where also the problem of reusing contextual knowledge is addressed.

Generally, the use of contextual information for pattern classification allows, by utilizing extra information, to reduce the ambiguity generated when identification of an object is inferred solely from its features.

2.2 Multiple classifier system

The majority of the works on the integration of context information fusion in classification leads to use multiple classifier systems. Indeed, those systems reported in the pattern recognition, statistics and machine learning literature have strong parallels with research on data fusion systems.

A multi-classifier systems (MCS) is a combination of classifiers from heterogeneous or homogeneous modeling backgrounds to give the final decision. The applications for such a system are numerous, where data available from multiple sources (or multiple sensors) generated by the same application may contain complementary information. Examples of application can be remote sensing data, computer security, financial risk assessment, fraud detection, recommender systems, and medical computer aided diagnosis.

The principle of multiple classifiers is that the diversity in the classifiers allows different decision boundaries to be generated by using slightly different training parameters, such as different training datasets. The intuition is that each classifier will make a different error, and strategically combining these classifiers can reduce total error [22]. Ensemble systems have attracted a great deal of attention over the last decade due to their reported superiority over single classifier systems on a variety of applications. They are nowadays highlighted by review articles which include extensive presentations of MCS concepts and architectures [23], [24].

In [22] an incremental learning algorithm Learn++ is reviewed and suitably modified for data fusion applications. The algorithm sequentially learns from data comprised of different sets of features by generating an ensemble of classifiers for each dataset, and then combining them through a modified weighted majority voting scheme.

Since the best combination of a set of classifiers depends on the application and on the classifiers, there is no single, best combination scheme nor any unequivocal relationship between the accuracy of a multiple classifier system and the individual constituent classifiers [25]. We define three categories for combining classifiers:

- Combination on different feature spaces, where a set of classifiers is combined each one designed on different feature spaces (perhaps using data from different sensors).
- Common feature space, where each classifier is defined on the same feature space and the combiner attempts to obtain a 'better' classifier through combination
- The final category of combination systems arise due to different classification of an object through repeated measurements. This may occur when we have a classifier designed on a feature space giving an estimate of the posterior probabilities of class membership, but in practice several (correlated) measurements may be made on the object.

Finally, the choice of combination depend generally on the real studied situation

3 Railway Information System with Context Information

We have applied the context information in a mixed-initiative spoken dialog system providing railway information system using spontaneous speech in Spanish [26]. A spoken dialog system (SDS) can be defined as a computer program that receives as input speech and generate as output synthesized speech, engaging the user in a dialog that aims to be similar to that between humans [27] [28]. Thus, these interfaces make technologies more usable, as they ease interaction [29], allow integration in different environments [28], and make technologies more accessible, especially for disabled people [30]

Usually, SDSs carry out five main tasks: Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Dialog Management (DM), Natural Language Generation (NLG), and Text-To-Speech Synthesis (TTS). These tasks are typically implemented in different modules of the system's architecture.

The goal of speech recognition is to obtain the sequence of words uttered by a speaker [31]. It is a very complex task, as there can be a great deal of variation in the input the recognizer must analyze, for example, in terms of the linguistics of the utterance, inter and intra speaker variation, the interaction context and the transmission channel. Once the speech recognizer has provided an output, the system must understand what the user said. The goal of spoken language understanding is to obtain the semantics from the recognized sentence. This process generally requires morphological, lexical, syntactical, semantic, discourse and pragmatic knowledge [32].

The dialog manager decides the next action of the system [26], interpreting the incoming semantic representation of the user input in the context of the dialog. In addition, it resolves ellipsis and anaphora, evaluates the relevance and completeness of user requests, identifies and recovers from recognition and understanding errors, retrieves information from data repositories, and decides about the next system's response. Natural language generation is the process of obtaining sentences in natural language from the non-linguistic, internal representation of information handled by the dialog system [33]. Finally, the TTS module transforms the generated sentences into synthesized speech [34]

As in many other spoken dialog systems, the semantic representation chosen for dialog acts of the SLU module of the railway information system is based on the concept of frame [35]. This way, one or more concepts represent the intention of the utterance, and a sequence of attribute-value pairs contains the information about the values given by the user. For the task, we defined eight concepts and ten attributes. The eight concepts are divided into two groups:

- Task-dependent concepts: they represent the concepts the user can ask for (*Timetables*, *Fares*, *Train-Type*, *Trip-Time*, and *Services*).
- Task-independent concepts: they represent typical interactions in a dialog (*Acceptance*, *Rejection*, and *Not-Understood*).

The attributes are: *Origin*, *Destination*, *Departure-Date*, *Arrival-Date*, *Ticket-Class*, *Departure-Hour*, *Arrival-Hour*, Train-*Type, Order-Number*, and *Services*. Figure 1 shows an example of the semantic interpretation of an input sentence.

Input sentence
Yes, I would like to know the timetables for tomorrow evening leaving from Valencia.
Semantic interpretation
(*Acceptance*) *Origin*: Valencia *Departure-Date*: Tomorrow

Departure-Hour: Evening

Figure 1. An example of the labeling of a user turn

The spoken dialog system considers the concepts and values for the attributes provided by the user throughout the previous history of the dialog to select the next system response. For the conversational agent to take this decision, we have assumed that the exact values of the attributes are not significant. They are important for accessing databases and for constructing the output sentences of the system. However, the only information necessary to predict the next action by the system is the presence or absence of concepts and attributes. Therefore, the codification we use for each concept and attribute is in terms of three values, {0; 1; 2}, according to the following criteria:

- (0): The concept is unknown or the value of the attribute is not given;
- (1): The concept or attribute is known with a confidence score that is higher than a given threshold;
- (2): The concept or attribute has a confidence score that is lower than the given threshold.

Using the previously described codification for the concepts and attributes, when a dialog starts (in the greeting turn) all the values are initialized to "0". The information provided by the users in each dialog turn is employed to update the previous values and obtain the current ones, as Figure 3 shows.

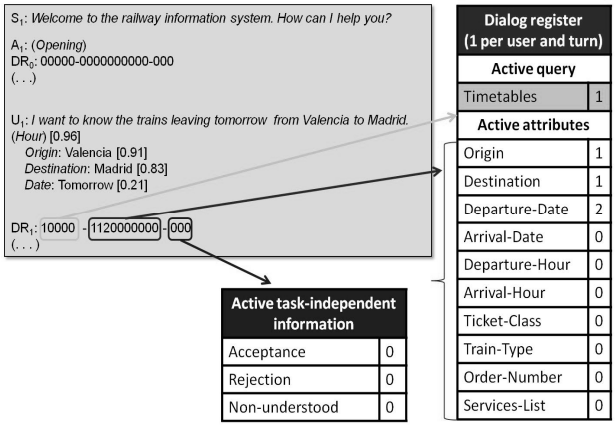

Figure 2. Excerpt of a dialog with its correspondent representation of the task-dependent and active task-independent information for one of the dialog turns

This figure shows the semantic interpretation and confidence scores (in brackets) for a user's utterance provided by the SLU module. In this case, the confidence score assigned to the attribute Date is very low. Thus, a "2" value is added in the corresponding position for this attribute. The concept (*Hour*) and the attribute *Destination* are recognized with a high confidence score, adding a "1" value in the corresponding positions.

Taking into account the codification of the current state of the dialog, which is denoted in Figure 3 by means of DR_1: 10000 – 1120000000-000, we propose the use of the previously described classification process that receives this codification as input and provides the probabilities of selecting each one of the system responses as output.

4 Experiments

The classification function can be defined in several ways. We have evaluated four different definitions of such a function: a multinomial naive Bayes classifier, n-gram based classifiers, a classifier based on grammatical inference techniques and a classifier based on neural networks [36]. The best results were obtained using a multilayer perceptron (MLP) [37], where the input layer holds the codification of the dialog state. The values of the output layer can be seen as an approximation of the a posteriori probability of belonging to the class associated to each one of the system responses.

A total of 51 system responses, i.e., system classes, were defined for the task (classified into confirmations of concepts and attributes, questions to require data from the user, and answers obtained after a query to the database).

An initial corpus of 900 dialogs (10.8 hours) was acquired for the task by means of the Wizard of Oz technique with 225 real users, for which an initial dialog strategy was defined by experts. A set of 20 scenarios was used to carry out the acquisition. Each scenario defined one or two objectives to be completed by the user and the set of attributes that they must provide. The corpus consists of 6,280 user turns, with an average number of 7.7 words per turn. The corpus was split into a training subset of 4,928 samples (80% of the corpus) and a test subset of 1,232 samples (20% of the corpus).

We defined three measures to compare the response automatically generated by the DM for each input in the test partition with regard to the reference answer annotated in the training corpus (the answer provided by the WOz). This way, the evaluation is carried out turn by turn. These three measures are:

- *exact*: the percentage of answers provided by the DM that are exactly the same that the reference response annotated in the training corpus;
- *strategy*: the percentage of answers provided by the DM that exactly follow the strategy defined for the WOz to acquire the training corpus;
- *coherent*: the percentage of answers provided by the DM that are coherent with the current state of the dialog although they do not follow the original strategy defined for the WOz.

- *error*: the percentage of answers provided by the DM that would cause the failure of the dialog;

Firstly we evaluated our proposal by carrying out a 5-fold cross validation process that considers only the semantic information provided by the SLU module for each user utterance, without any additional context information. Table 1 shows the results of this evaluation.

exact	*strategy*	*coherent*	*error*
78.62%	97.73%	99.68%	0.18%

Table 1. Evaluation of the proposed statistical methodology

Secondly, we evaluated our proposal considering the gender of the users as a context parameter. To do this, the corpus was divided into a set of partitions with equal number of samples of women and men. Table 2 shows the results of this experimentation, indicating the partitions used for training and test (Training/Test).

	exact	*strategy*	*coherent*	*error*
Women / Both	70,55%	92,56%	96,93%	0,73%
Men / Both	69,97%	95,92%	97,96%	0,58%
Both / Women	71,72%	94,67%	97,75%	0,49%
Both / Men	76,09%	97,79%	99,71%	0,14%

Table 2. Evaluation of the proposed statistical methodology taking into account the influence of gender

The results of this evaluation show that there are not remarkable differences in the model if the learning of the model is made solely using samples of men or women (first two columns of results in Table 2). Higher differences are observed in the evaluation of the model considering the gender of the users in the test partitions (third and fourth column of this table). The differences obtained in these cases indicate a greater similarity in the samples of men.

Thirdly, we have evaluated our proposal taking into account the influence of the origin of the dialogs. This evaluation starting with the same partitions defined for the complete evaluation of the proposal, training the dialog model with the samples coming from the specific location to be evaluated and using the same test partitions (samples from the three possible locations). Table 3 shows the results of this evaluation.

	exact	*strategy*	*coherent*	*error*
Location 1	71,94%	95,55%	99,16%	0,42%
Location 2	77,24%	95,51%	97,11%	1,44%
Location 3	70,39%	90,03%	92,15%	3,92%

Table 3. Evaluation of the proposed statistical methodology taking into account the origin of the dialogs

The results of this evaluation show the better operation of the methodology when the dialog model was learned with the dialogs acquired at Location 1. Learning the model only with the dialogs of the Location 2, the percentage of responses that follow the strategy is equivalent to the one obtained for the Location 1. Nevertheless, the percentage of erroneous answers (*error*) is three times greater. With regard the dialog corpus acquired at Location 3, in spite of obtaining a percentage of correct answers of 92.15%, the number of system answers that can cause the failure of the dialog is the highest. Therefore, considering the *error* measurement, we can conclude a significant difference between the dialogs acquired at each location.

5 Conclusions

In this work, we propose the use of context information to improve the performance of automatic dialog system in railway information system. The main idea is to incorporate information of the context of the sequences of phrases to improve the classification task.
Future work could use a set of classifiers to develop a multi-classifier system able to use the context information to adapt the fusion process or to select the most appropriate one.

Acknowledgement

This work was supported in part by Projects MINECO TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485)

References

[1] D. Hall et J. Llinas, «An introduction to multisensor data fusion,» *Proceedings of the IEEE,* vol. 85, 11, pp. 6-23, 1997.

[2] J. Llinas, «A Survey and Analysis of Frameworks and Framework Issues for Information Fusion Applications,» chez *Hybrid Artificial Intelligence Systems*, Madrid, Springer Berlin Heidelberg, 2010, pp. 14-23.

[3] B. Vijay et J. Wilson, «Survey of Context Information Fusion for Sensor Networks based Ubiquitous Systems,» *Journal of Sensor and actuator networks,* 2013.

[4] E. F. Nakamura, A. A. F. Loureiro et A. C. Frery, «Information fusion for wireless sensor networks: Methods, models, and classifications.,» *ACM Comput. Surv. ,* vol. 39, 2007.

[5] A. N. Steinberg, C. L. Bowman et F. E. White, « Revisions to the JDL data fusion model,» *Proc. SPIE 3719, Sensor Fusion: Architectures, Algorithms, and Applications III, ,* p. 430, 1999.

[6] L. Snidaro, J. Garcia et J. M. Corchado, «Context-based information fusion,» *Information Fusion Guest Editorial.,* n 121 , p. 82–84, 2015.

[7] M. A. Patricio, J. García, J. M. Corchado, J. Bajo, A. Khamis et E. E. Mangina, «Intelligent Systems in Context-Based Distributed Information Fusion,» *International Journal of Distributed Sensor Networks,* 2013.

[8] L. Snidaro, J. García et J. Llinas, «Context-based Information Fusion: A survey and discussion,» *Information Fusion,* 2015.

[9] J. Gómez-Romero, J. García, M. Kandefer, J. Llinas, J. Molina, M. Patricio, M. Prentice et S. Shapiro, «Strategies and Techniques for Use and Exploitation of Contextual Information in High-Level Fusion Architectures,» *Proceedings of the 13th International Conference on Information Fusion,* p. 2010.

[10] M. Kandefer et S. C. Shapiro, «Evaluating spreading activation for soft information fusion,» *Proceedings of the 14th International Conference on Information Fusion, IEEE,* 2011.

[11] J. McCarthy, «Notes on formalizing context,» *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI93), IEEE, Chambery, France,* p. 555–560, 1993.

[12] G. L. Rogova, «Context-awareness in crisis management,» *Proceedings of the Military Communications Conference, IEEE,* pp. 1-7, 2009.

[13] E. G. Little et G. L. Rogova, «Designing ontologies for higher level fusion,» *Information Fusion,* vol. 10, p. 70–82, 2009.

[14] J. Gòmez-Romero, M. A. Serrano, J. Garcìa, J. M. Molina et G. Rogova, «Context-based multi-level information fusion for harbor surveillance,» *Information Fusion,* vol. 21, p. 173–186, 2015.

[15] A. N. Steinberg, C. L. Bowman, G. Haith et E. Blasch, «Adaptive context assessment and context management,» *Proceedings of the International conference on Information Fusion,* 2014.

[16] A. N. Steinberg et C. L. Bowman, «Adaptive context discovery and exploitation.,» *Proceedings of the16th International Conference on InformationFusion,* 2013.

[17] A. Makarau, G. Palubinskas et P. Reinartz, «Multi-sensor data fusion for urban area classification,» *Joint Urban Remote Sensing Event,* 2011.

[18] H.-H. Wang, Y.-S. Lu et M.-J. Chen, «Multisensor Information Fusion Application to SAR Data Classification,» *Springer-Verlag Berlin Heidelberg,* p. 364 – 373, 2006.

[19] X. B. Song, Y. Abu-Mostafa, J. Sill, H. Kasdan et M. Pavel, «Robust image recognition by fusion of contextual information,» *Information Fusion ,* vol. 4, n 13, p. 277–287, 2002.

[20] O. Brdiczka, P. C. Yuen, S. Zaidenberg, P. Reignier et J. L. Crowley, «Automatic acquisition of context models and its application to video surveillance,» *18th International Conference on Pattern Recognition, IEEE,* vol. 1, pp. 1175-1178, 2006.

[21] L. Snidaro, I. Visentini, J. Llinas et G. L. Foresti, «Context in fusion: some considerations in a JDL perspective,» *Proceedings of the 16th International Conference on Information Fusion, IEEE,* 2013.

[22] R. Polikar, L. Udpa, S. Udpa et V. Honavar, «Multiple Classifier Systems for Multisensor Data Fusion,» *SAS 2006 – IEEE Sensors Applications Symposium,* 2006.

[23] M. Wozniak, M. Graña et E. Corchado, «A survey of multiple classifier systems as hybrid systems,» *Informat. Fusion,* 2013.

[24] L. Rokach, «Taxonomy for characterizing ensemble methods in classification tasks: a reviw and annotated bibliography,» *Computational statistics and data analysis,* vol. 12, pp. 4046-4072, 2009.

[25] Kuncheva LI & Whitaker CJ: *Feature Subsets for Classifier Combination: An Enumerative Experiment.* Lecture Notes in Computer Science, vol 2096, Springer-Verlag, Berlin

[26] D. Griol, L. Hurtado, E. Segarra et E. Sanchis, «A Statistical Approach to Spoken Dialog Systems Design and Evaluation,» *Speech Communication 50,* p. 666–682, 2008.

[27] R. Pieraccini, «The Voice in the Machine: Building Computers that Understand Speech,» *The MIT Press ,* 2012.

[28] T. Heinroth et W. Minker, «Introducing Spoken Dialogue Systems into Intelligent Environments,» *Kluwer Academic Publishers Springer-Verlag.,* 2010.

[29] T. Hempel, «Usability of Speech Dialog Systems: Listening to the Target Audience,» *Springer,* 2008.

[30] R. Vipperla, M. Wolters et S. Renals, «Spoken dialogue interfaces for older people,» *Advances in Home Care Technologies ,* p. 118–137, 2012.

[31] A. Tsilfidis, I. Mporas, J. Mourjopoulos et N. Fakotakis, «Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing,» *Computer Speech & Language,* vol. 27 , p. 380–395, 2013.

[32] W. L. R.-Z. Wu, J.-Y. Duan, H. Liu, F. Gao et Y. Chen, «Spoken language understanding using weakly supervised learning,» *Computer Speech & Language,* vol. 24, p. 358–382, 2010.

[33] V. López, E. Eisman, J. Castro et J. Zurita, «A case based reasoning model for multilingual language generation in dialogues,» *Expert Systems with Applications,* n° 139, p. 7330–7337, 2011.

[34] T. Dutoit, «An Introduction to Text-To-Speech Synthesis,» *Kluwer Academic Publishers,* 1996.

[35] M. Minsky, «A Framework for Representing Knowledge,» *The Psychology of Com-puter Vision. McGraw-Hill ,* p. 211–277, 1975.

[36] G. David, C. Zoraida, L.-C. Ramón et R. Giuseppe, «A domain independent statistical methodology for dialog management in spoken dialog systems,» *Computer Speech & Language 28(3),* pp. 743-768 , 2014.

[37] D. E. Rumelhart, G. E. Hinton et R. J. Williams, «PDP: Computational models of cognition and perception,» *MIT Press, Ch. Learning internal representations by error propagation,* pp. 319-362, 1986.