PEBM: A Probabilistic Exemplar Based Model

Andres F. M. Rodriguez
Institute) de Investigaciones Electricas
Cuernavaca
Mexico
afrm@iie.org.iax

Sunil Vadera
School of Sciences
University of Salford
Salford M5 4WT, UK
S.Vadera@cms.salford.ac.uk

Abstract

A central problem in case based reasoning (CBR) is how to store and retrieve cases. One approach to this problem is to use exemplar based models, where only the prototypical cases are stored. However, the development of an exemplar based model (EBM) requires the solution of several problems: (i) how can a EBM be represented? (ii) given a new case, how can a suitable exemplar be retrieved? (iii) what makes a good exemplar? (iv) how can an EBM be learned incrementally? This paper develops a new model, called a probabilistic exemplar based model, that addresses these questions. The model utilizes Bayesian networks to develop a suitable representation and uses probabilistic propagation for assessing and retrieving exemplars when a new case is presented. The model learns incrementally by revising the exemplars retained and by updating the conditional probabilities required by the Bayesian network. The paper also presents the results of evaluating the model on three datasets.

1 Introduction

Case Based Reasoning (CBR) is an approach that utilises past situations in an attempt to solve new problems. The basic CBR cycle involves retrieving cases that are similar to the current problem and utilising them to solve the current problem. This makes memory organisation and indexing a fundamental part of CBR systems. One approach is to store a flat database of cases and scan all the cases to identify the most similar cases. For applications where many more are involved, this simple organisation is considered to be slow [Kolodner, 1993].

A more sophisticated method is to partition the cases into clusters and organise them hierarchically. The hierarchy can then be searched more efficiently by following a path depending on the features of the new case. Different types of hierarchies have been proposed leading to different approaches. One approach is to use decision trees so that the leaf nodes contain the cases and where the internal nodes contain questions that can be used to partition the cases. So for example, systems like ReMind [Althoff *et* o/., 1995] provide a tree induction algorithm that can be used to avoid examining all the cases. This kind of approach is particularly useful when large databases of cases are already available. However, when cases are not available in advance, and the domain is not well defined this approach is more difficult to apply.

Another approach is to use an abstraction hierarchy where each internal node is an abstraction of the cases represented by its children. These hierarchies are known as discrimination networks or redundant discrimination networks when the nodes represent overlapping regions of cases. A number of research systems, such as MEDIATOR and CASEY (see [Kolodner, 1993] for a description and references) have used this approach and their outcomes have shown its utility. However, these systems require much more memory to store the network and the procedures for adding new cases arc very expensive since the abstraction hierarchy may need to be restructured [Kolodner, 1993].

Thus, current approaches to CBR work well in some situations, but also have problems in other situations. In particular, for domains, sometimes called *weak dojnains* [Porter *et* a/., 1990], where: (i) the categories or concepts are difficult to define by necessary and sufficient features, (ii) the categories can be non-disjoint, (iii) the data are not structured, (iv) all the data do not exist in advance, and (v) there is uncertainty in how the categories are represented by cases, these approaches have limitations since they often require all the features and examples in advance, and do not handle uncertainty explicitly (they use a weighted sum of the differences).

An alternative approach, that is perhaps more applicable to weak domains, is to store only prototypical cases. This approach, known as the *exemplar" based* model has its basis in cognitive theories, which postulate that concepts can be represented by exemplars [Smith and Medin, 1981]. Exemplar based models do not necessarily require all the features or all the cases in advance.

This paper focuses on developing an exemplar based model. The next section presents the main problems of developing exemplar based models. The paper is organised as follows: section 3 presents the model in terms

of the knowledge representation, the classification, and learning processes; section 4 presents an empirical evaluations on three datasets; and section 5 presents the conclusions.

2 The Problem

To understand the problem, consider the diagram shown in Fig. 1 that shows a weak domain in which there are two categories *A* and *B* (solid lines). The category *A* has nine cases (the points) $c_1, c_2, c_3, c_4, c_6, c_7, c_8, c_9$, and c_{10}, and the category *B* has five cases c_3, c_4, c_5, c_9, and c_{11}. Note that the cases c_3, c_4 and c_9 are common cases that occur in both categories. The main problem is to

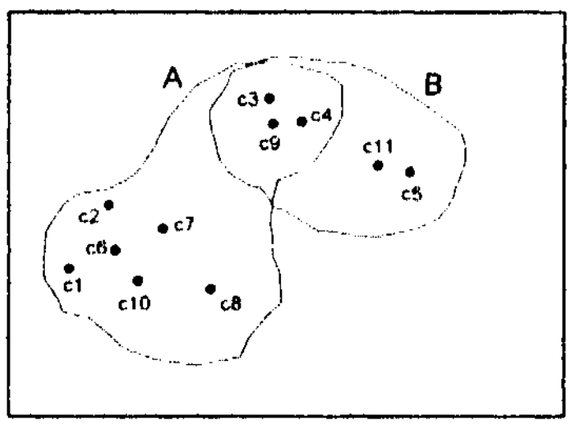

Figure 1: Example of a weak domain.

proceed from a view like the one shown in Fig. 1 to an exemplar based view like the one shown in Fig. 2 where the exemplars e_6, e_8, e_9, and e_{11} represent sets of similar cases (dashed lines). That is, instead of storing all the

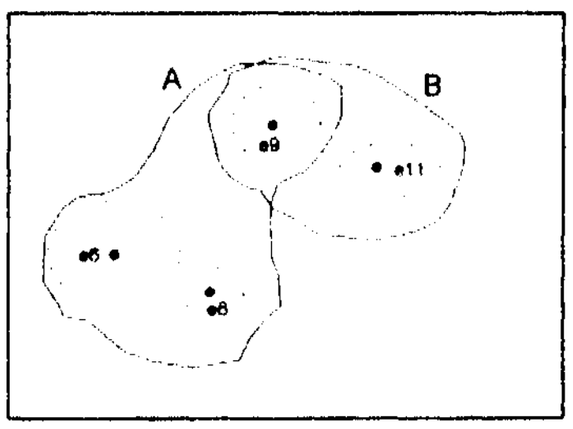

Figure 2: Exemplar based view in weak domain.

cases, only the prototypical cases are stored. Although conceptually, this is an elegant idea, attempting to develop it raises the following difficult questions:

1. What is a good representation of the model?
2. How can a new case be classified?
3. What notion of similarity can be adopted?
4. What makes a good exemplar?
5. How can the model be learned incrementally?

The next section of this paper develops the model by addressing these questions.

3 The Model

3.1 The Knowledge Representation

One way of representing the information in Fig. 2 is to use a network in which nodes can be used to denote exemplars, features, and categories. Thus, Fig. 3 shows the network representing the exemplar based model shown in Fig. 2. In this representation, the dashed lines show the relationship between categories and exemplars, and the solid lines show the relationship between exemplars and their features. So for example, category *A* has the exemplars e_6, e_8, and e_9 and exemplar e_6 has the features f_1,f_2, and F_3. Notice that exemplars can be shared by categories, and features can be shared by exemplars.

As it stands, Fig. 3 is not an adequate representation of an exemplar based model since it does not contain any information about the degree of dependency between a category and its exemplars and an exemplar and its features. So for example, a car can have features such as colour, engine, and make. But, which of them is more relevant in the representation of a car? The above representation would not differentiate between the strong dependency: an object being a car and having an engine, and the weak dependency: an object being a car and its colour.

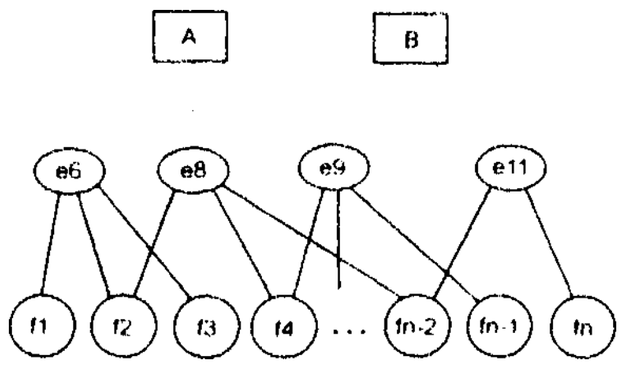

Figure 3: A basic exemplar based representation.

Hence, to include the strengths of such dependencies, the relationships between exemplars and features are represented as probabilistic dependencies. That is, each feature f_j, that is a leaf node in the network, is labelled with the conditional probability $P(f_j \mid e_1 \cdots e_k)$, where $e_1 \cdots e_k$ are the exemplars that share the feature f_j. Similarly, the importance of an exemplar in the category is represented by probabilistic dependencies. Each exemplar c,, which is an intermediate node in the network, is labelled with the conditional probability $P(e_i \mid JC)$, where *JC* is the joint category formed by the parents of e_i. This probability is the prior probability of the exemplar when no evidence is available. With this additional information, the network of Fig. 3 becomes a hybrid representation.

3.2 The Classification Process

Given the above representation, how can the following questions, raised earlier, be addressed:

- How can a new case be classified?
- What notion of similarity can be adopted?

The majority of current CBR systems address these questions by adopting a similarity metric, which is a weighted sum of the differences between a new case and a stored case [Kolodner, 1993]. The main problem with this approach is that the weights of the similarity metric need to be estimated and obtaining reliable weights is not easy for non-triviai problems (see [Wettschereck *et al,* 1997] for a survey).

In this paper, the notion of similarity adopted is that two cases are similar if they are represented by the same exemplar. But how can one determine if a new case is represented by a particular exemplar? Since in the above representation, the lower network that relates exemplars and features is a Bayesian network, the degree to which a new case with features f_{inc}-, f_{qnc} is represented by an exemplar *e* can be computed by:

$$P(e \mid f_{1nc}, \cdots, f_{qnc})$$

This computation can be carried out by using propagation methods developed by Pearl [1988].

Given this capability of calculating the extent to which an exemplar represents a new case, all the exemplars could be investigated, in theory at least. However, probabilistic propagations methods can be computationally expensive (it is known to be NP-hard in general) and investigating all the exemplars is therefore not practical.

Hence, first it is necessary to rank the categories in order of the likelihood of them containing a suitable exemplar. This ranking has to be performed in a way that avoids missing suitable exemplars but is computationally efficient. This ranking can be obtained by utilizing an observation by Smith and Medin [1981] who point out that:

> "the features that represent a concept are salient ones that have a substantial probability of occurring in instances of the concept".

Thus, the important features will have high values of occurrence given an exemplar, i.e., high values of *P(fj | e).* Hence, a reasonable way of ranking the categories is to obtain the contribution of the features of the exemplar that are present in the new case, averaged over the number of features in the exemplar e_i

$$Rank(e_i) = \frac{\sum_{f \in e_i} P(f \mid e_i)}{nfe_i}$$

where

$$P(f \mid e_i) = 0 \quad when \ f \notin nc$$

In this equation, *nc* is a new case and n/e_i is the number of features in the exemplar e_i.

Then, the categories can be ranked in order of the rank of their exemplars. Once the ranking is obtained, a suitable investigation strategy can be adopted. For example, the list of categories can be investigated in order of rank until a good exemplar is found. Within each category, propagation methods can be used to assess the merits of an exemplar *e* by computing $P(e \mid nc)$ and stopping if this is above a threshold that is normally dependent on the application.

3.3 The Learning Process

Learning an exemplar based model incrementally involves two aspects: (i) learning the the model and (ii) estimating its parameters, both of which are described in this subsection.

Learning the model

The learning process of an exemplar based model needs to answer the following questions, that were raised earlier:

1. What makes a good exemplar?
2. How can the model be learned incrementally?

To answer these questions, consider a situation where there is a category *C* that is represented by three exemplars e1,e<2 arrives- Suppose a new training case with category *C* arrive can arise: (i) the new case is not classified by the exemplars in C, and (ii) the new case is correctly classified by an exemplar in *C.*

In the first case, clearly the new case should be retained as a new exemplar since it must be different from the other exemplars. In the second case, criteria need to be developed for deciding which of the two, the new case or exemplar, will be the best representative of all cases in the region.

For exemplar based models these criteria have to be based on the notion of *prototypicality.* Before describing the measure of prototypicality used in this work, it is necessary to first describe the idea of a summary representation. Earlier, an exemplar was represented as a Bayesian network with dependencies from the exemplar to its features. In general, an exemplar may not have the same features as all the similar cases that it represents. For example, an exemplar e_2 may have the features f_4, f_6, and f_9 while the union of all the features of the cases it represents may be f_3,f_4,f_6,f_7, and f_9. A *summary representation* is a Bayesian network where all the features of the similar cases are included.

Returning to the notion of prototypicality, the problem is to develop a measure of prototypicality so that a good prototype can be selected. Rosch and Mervis [1975] argued that a case is an ideal prototype if :

- it has the highest family resemblance with other members in the same category, (this is known as *focality* [Biberman, 1995]) and
- it has the least family resemblance with members of other categories (this is known as *peripherality* [Biberman, 1995]).

In the context of the model being developed here, *family resemblance* is viewed as the collection of similar cases and which have a summary representation. In terms of regions, a case that maximizes the probability of covering a region can be considered to have the highest family resemblance. Since the summary representation denotes regions, and takes the form of a Bayesian network, a suitable measure of focality of an exemplar e_i is the probability of covering a region:

$$Focality(e_i) = P(SR(e_i) \mid e_i)$$

where *SR(ei)* denotes the summary representation of the region that contains e_i.
Likewise, a suitable measure of peripherally is obtained by working out the average probability of an exemplar representing regions in other categories:

$$Peripherality(e_i, C) = \frac{1}{k}\sum_{j=1}^{k} P(SR(e_j) \mid e_i) \;\; \forall j \neq i \in C$$

These two measures can be used to define a measure of prototypicality as follows. Since a good prototype is one that has the greatest focality and the least peripherally, the measure of prototypicality adopted here is:

$$Prototypical(e_i, C) = Focality(e_i) - Peripherality(e_i, C)$$

This measure of prototypicality can now be used to decide which case makes the better exemplar in a region. These considerations lead to a learning algorithm that can be summarised as follows. Given a new training case, *nc,* first use the classification process described above. If the case is not classified successfully, then the training case becomes an exemplar. If it is correctly classified by an exemplar *e* then use the above prototypicality measure to determine which of the two best represents the region of cases and retain the more representative one.

Estimating the probabilities

To use the above classification and learning processes, one needs the probabilities that define the Bayesian network which represents the exemplars. Since the model is incremental, and the cases are not retained, estimating the probabilities in a manner that enables a good exemplar based model to be learned is a non-trivial problem.

The Bayesian exemplar based model requires the estimation of two parameters that need to evolve as new cases are seen:

1. prior probabilities of the exemplars in the joint category $P(e \mid C)$ and
2. the conditional probability $P(f \mid parents(f))$.

The first of these is obtained in a standard way by utilising the beta distribution which leads to the following equation (see [Lindgren, 1976] for details) that can be used to compute and update the prior probabilities:

$$P(e \mid C) = \frac{number\ of\ cases\ repr.\ by\ e + 1}{number\ of\ cases\ in\ C + 2}$$

Estimating the conditional probabilities *P(f\parents(f))* is much more difficult. In general, 2^{n+1} probabilities need to be estimated for *n* parents. In particular, there may not be enough cases in the intersection of the parent events, even if there are enough cases in the regions represented by the parents. This means that estimates of probabilities such as $P(f \mid \neg e1, e2)$ could only be based on a small number of cases and would therefore be inaccurate even when many cases have been seen.

To overcome this problem, the noisy or model [Pearl, 1988] is considered. If this model can be adopted, then instead of requiring $P(f|parents(f))$ only $P(f \mid e_i)$ is needed, for each parent e_i of f. To see if the noisy or model can be used, consider the assumptions that it makes [Pearl, 1988]:

Accountability An event m_j is false, $P(m_j)$ = 0, if all conditions listed as causes of m_j are false.

Exception independence If an event m_j is a consequence of two conditions d_1 and d_2, then the inhibition of the occurrence of m_j under d_1 is independent of the mechanisms of inhibition of m_j under d_2.

In the context of this model, the exception independence assumption can be interpreted as requiring that the absence of the feature given one exemplar is independent of the absence of the feature given another exemplar. The extent to which this assumption holds depends on the way the exemplars are selected. In Section 3.3, the selection scheme uses a measure of prototypicality that aims to reduce the possibility of selecting exemplars that represents similar regions. That is, the selection scheme used minimizes the possibility of the exception independence assumption being broken.

The accountability assumption requires that if a case is not represented by the parent exemplars of a feature, then that feature does not occur in the case. Although this may hold when an accurate exemplar based model has been learned, it clearly does not hold while it is still learning (e.g. consider a new case that should be a new exemplar). To overcome this problem, an additional virtual exemplar is added in the representation of each category. This additional exemplar can be viewed as representing all the cases that have not yet been seen and therefore ensures that the accountability condition holds. With this additional exemplar, the revised model is illustrated in Fig. 4. As the figure shows, this introduces dependencies between the virtual exemplar and the features. But how can the strengths of the dependencies be estimated, since the virtual exemplar represents unseen cases? Estimating the strengths of these dependencies is

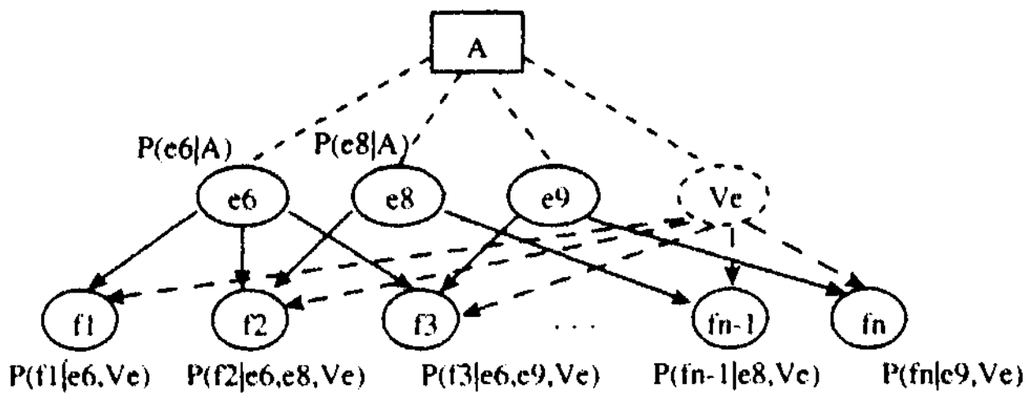

Figure 4: Virtual exemplar.

a task that requires predicting the behaviour of the dependencies as more cases are observed. This behaviour can be expected to have the following characteristics:

- The strengths of the dependencies should be the highest initially when no cases have been seen and ignorance is greatest.
- As more cases are observed, the strengths of the dependencies can be expected to decay since the

virtual exemplar will represent fewer unseen cases.

- There is always a small chance that a new case will be in the region represented by the virtual exemplar even after many cases have been observed.

There may be several functions that satisfy these characteristics. However, this work utilises the exponential function, which is often used to represent decay (e.g., in modelling radioactivity) and takes the form:

$$P(f \mid Ve) = \lambda e^{-\lambda * a * n} \quad OR \;\; 0.1 \; if \;\; P(f \mid Ve) < 0.1$$

where *n* is the number of cases in a category and the parameters *a* and A determines the rate of decay. The lower bound of 0.1 in this function reflects the possibility that a new case will be in the region represented by the virtual exemplar even after many cases have been seen.

This completes the description of how the probabilities can be learned incrementally, thereby allowing the use of the classification and the learning procedures.

4 Empirical Evaluation

The model described has been implemented and evaluated on three datasets, votes, zoo, and audiology, available from the University of California repository of datasets. Due to the lack of space, this section presents only a very brief summary of the results which are presented more fully in [Rodriguez, 1998]. Table 1 summarises the characteristics of each dataset. Each experiment randomly partitioned the data into a 70% training set and a 30% testing set, and was repeated 20 times to obtain an average accuracy and a compression ratio (defined as the proportion of cases not retained) for each class. All the experiments were carried out with the parameters A and *a* set to 0.6 *v* d 0.1 and a threshold of 0.75. Although no attempt, has been made to find optimal values for these parameters, the model works well with these values. The problem of recommending their optimal values given the characteristics of the domain is a subject for future work. Tables 2 and 3 give the results for the votes and the zoo datasets. As the results show, the model performs well both in terms of accuracy and the number of exemplars retained. The overall compression ratio for the votes dataset is 97% with an accuracy of 89%. The overall compression ratio for the zoo dataset is 87% with an accuracy of 92%.
In Table 3, an interesting difference in accuracy occurs between class-3, which has a low accuracy of 16%) and class-5 which has an accuracy of 77% and both classes have about 3 training cases on average. A close look at the classes reveals the reason for this behaviour. Class-3 consists of five relatively different animals: pitviper, seasnake, slowworrn, tortoise, and tuatara, while class-5 consists of fairly similar animals: frog, poisonous frog, newt, and toad. Since, class-3 is very polymorphic and only a few cases have been observed, the exemplars representing that category are weak and hence the accuracy of class-3 is low. However, although there are only a few cases in class-5, they are similar and the exemplars are therefore more representative of the category. Hence, the accuracy for class-5 is significantly better.

Figure 5 present, the accuracy obtained, and the number of training cases per category for the audiology dataset. The accuracies obtained are good for some of the categories and poor for some categories where there are few training cases. This behaviour is to be expected since the model is not expected to learn exemplars from a few cases. The motivation for applying PEBM to the audiology dataset was to enable some comparison with a closely related system, PROTOS [Bareiss, 1989] that utilised that data. Unfortunately, due partly to the nature of PROTOS, and partly because of the lack of availability of the information utilised in the PROTOS experiments, it is not possible to repeat the experiments reported in [Bareiss, 1989]. Although it would be incorrect to draw comparative conclusions from the results of the single trial presented in [Bareiss, 1989], it is encouraging that the results obtained in terms of accuracy and compression ratio, are similar to those obtained when PROTOS was trained with the aid of an audiology expert (see [Rodriguez, 1998] for details).

Table 1: A summary of the datasets.

Dataset name	No. of cases	No. of features	No. of concepts	Missing values
Votes	435	16	2	Y
Zoo	101	16	7	N
Audiology	226	69	23[1]	Y

Table 2: Averages results for the votes dataset.

Category	Training cases	Test cases	Exemplars	Accuracy ±95% CI
Republicans	119.20	47.80	2.1	96%±1.9%
Democrats	185.05	81.95	4.0	84%±2.0%

Table 3: Averages results for the zoo dataset.

Category	Training cases	Test cases	Exemplars	Accuracy ±95% CI
class-1	28.40	12.60	1.35	98%±1.7%
class-2	14.16	5.35	1.00	99%±1.4%
class-3	3.45	1.55	1.65	16%±12.3%
class-4	9.30	3.70	1.00	100%±0%
class-5	2.90	1.10	1.00	77%±19.3%
class-6	5.35	2.65	1.00	100%±9.8%
class-7	7.30	2.70	1.75	80%±9.6%

5 Related Work and Conclusion

The model presented in this paper is related to work on CBR, Bayesian networks, and inductive learning. There are numerous systems in these categories and a comparison with these systems is too lengthy for the space avail-

[1] Thc category of all the unclassified cases is omitted.

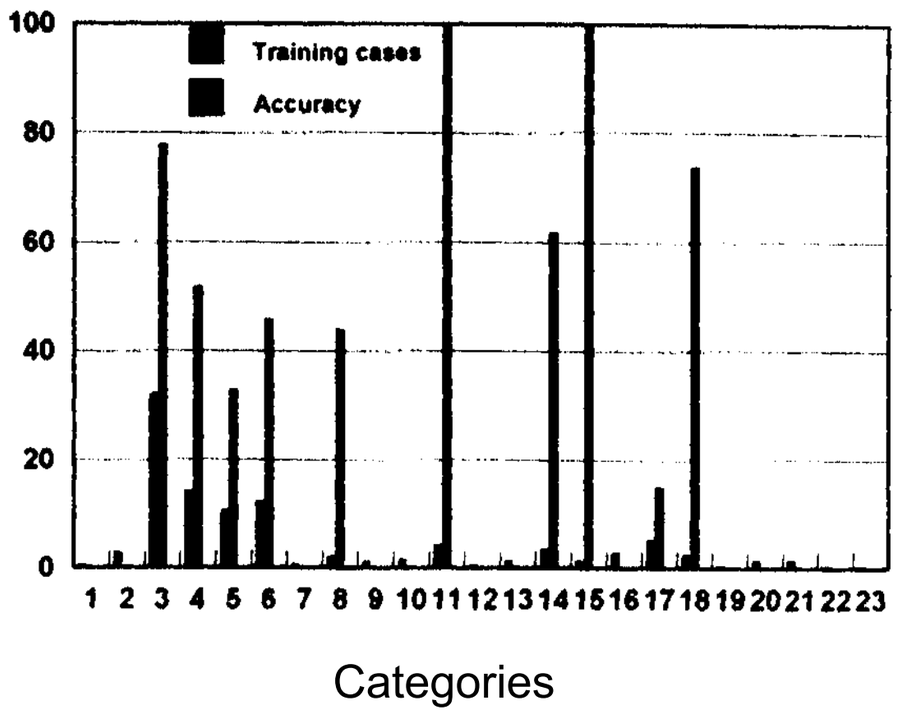

Figure 5: Accuracy for the audiology dataset.

able in this paper. It is, however, important to merit ion the main differences with two of the most related models: PROTOS and Tirri et al.'s [1996] model. The representation used by PROTOS is similar to the one used by PEBM in that exemplars are used to define categories. The notion of exemplar is, however, very different in that cases denote exemplars, whereas in PEBM, exemplars are represented by Bayesian networks. The classification process used by PROTOS is dependent on the use of indices called remindings, censors, and difference links. In contrast, classification in PEBM is achieved by probabilistic propagation. The learning mechanisms are also very different since PROTOS relies heavily on heuristics that learn from user provided explanations, while PEBM learns from data. The most significant difference, however, is that PEBM has foundations in probabilistic reasoning, whereas PROTOS appears to be based primarily on heuristics.

The Bayesian network representation used in Tirri et al.'s [1996] work is very similar to the one adopted for PEBM but with the exception that their upper level nodes are random variables that represent cases and not prototypes. Given the potentially large number of cases, standard propagation methods would not be practical. Hence, they assume that the cases are mutually exclusive in order to simplify the network to a tree. The extent to which this assumption holds or the effects of violating the assumption are unclear since a new case can be expected to be similar to a number of previous cases. In contrast, PEBM does not make this assumption and uses exemplars which aim to represent regions of similar cases. This difference is also reflected in the requirements for learning, since their model only estimates the probabilities from all the cases, while PEBM identifies prototypes incrementally.

To conclude, this paper has presented an exemplar based model with foundations in Bayesian networks. The model learns exemplars by using a measure of prototypicality and utilises probabilistic propagation to determine whether a new case is similar to an exemplar. The model has been evaluated on 3 datasets and shows promising results in terms of accuracy and in terms of the number of exemplars retained.

Acknowledgements

The authors are grateful to Enrique Sucar for his useful comments and discussions on the model presented in this paper.

References

[Althoff *et al,* 1995] K. Althoff, E. Auriol, R. Barletta, and M. Manago. A review of industrial case-based reasoning tools. *AI Intelligence,* 1995.

[Bareiss, 1989] R. Bareiss. *Exemplar-based knowledge acquisition. A unified approach to concept representation, classification, and learning.* Academic Press Inc., U.S.A., 1989.

[Biberman, 1995] Y. Biberman. The role of prototypieality in exemplar-based learning. In Nada Lavrac and Stefan Wrobel, editors, *Proc. of Machine Learning: ECML-95, 8th European Conference on Machine Learning,* pages 77 91, Heraclion, Crete, Greece, 1995.

[Kolodner, 1993] J. Kolodner. *Case-based reasoning.* Morgan Kaufmann, Palo Alto, CA, U.S.A., 1993.

[Lindgren, 1976] B.W. Lindgren. *Statistical Theory.* Macmillan Publishing Co. Inc., 1976.

[Pearl, 1988] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan kaufmann, Palo Alto, CA, U.S.A., 1988.

[Porter *et al,* 1990] B.W. Porter, R. Bareiss, and R.C. Holte. Concept learning and heuristic classification in weak-theory domains. *Artificial Intelligence,* 45:229 263, 1990. "

[Rodriguez, 1998] Andres F. Martinez Rodriguez. *A Probabilistic Exemplar Based Model.* PhD dissertation, Salford University, Salford, M5 4WT, 1998.

[Rosch and Mervis, 1975] E. Rosch and C.B. Mervis. Family resemblance studies in the internal structure of categories. *Cognitive Psychology,* 7:573-605, 1975.

[Smith and Medin, 1981] E. Smith and D. Medin. *Categories and concepts.* Harvard University Press, U.S.A., 1981.

[Tirri *et al,* 1996] H. Tirri, P. Kontkanen, and P. Myllymaki. A bayesian framework for case-based reasoning. In *Proc. of the 3rd European Workshop on Case-Based Reasoning,* pages 413 427, Switzerland, 1996.

[Wettschereck *et a*/., 1997] D. Wettschereck, D.W. Aha, and T. Mohri. A review and empirical evaluation of feature-weighting methods for a class of lazy learning algorithms. *AI Review,* 11:273 314, 1997.