Chapter 11
Discussion of Research Contributions

> 'In theory every individual scientist is capable of being his/her most severe critic, and his/her own writings are expected to discuss with real care and seriousness the objections against his/her own novel ideas..' (Toulmin et al, 1979)

The research contributions for feature selection, base model design and dataset selection for aggregate modeling were summarised in sections 10.2.1 and 10.3.1. It was stated in chapter 4 that the design science research paradigm was used to guide the activities of the research, and the design science research process was discussed in detail in that chapter. A brief discussion of the expected design science research outputs is provided in this chapter followed by the author's self-assessment of how the research meets the expectations of design science research. Sections 11.1 and 11.2 respectively provide a discussion of the outputs of design science research and the recommendations for design science research evaluation. Section 11.3 provides a discussion of the limitations of the methods proposed in this thesis. Section 11.4 provides a summary of this chapter.

11.1 Outputs of design science research

Hevner et al (2004) have stated that design science research for Information Systems must produce one or more artifacts. Recall from chapter 4 that Hevner et al (2004) have defined an artifact as:

> '..innovations that define ideas, practices, technical capabilities, and products, through which the analysis, design, implementation, and use of Information Systems can be effectively accomplished.'

Hevner et al (2004) and March and Smith (1995) have further stated that the artifacts for design science research are *constructs, models, methods*, and *instantiations*. Vaishnavi and Kuechler (2004/5) have observed that in addition to the production of artifacts, design science research should produce *better theories* for the field of research. *Constructs* form the conceptual vocabulary of the field of study. *Constructs*

make up the language used to define and communicate the problems and solutions in the field of study. For design science research, the term '*model*' is used to refer to the set of propositions that specify relationships between the *constructs*. *Methods* are definitions of the processes that need to be achieved. A method may be stated as a set of steps to perform a given task, or a method may be specified as a formal computational algorithm. *Instantiations* are the actual implementations of the models and methods in order to demonstrate that they actually work. '*Better theories'* provide an increased understanding arising from the study of the created artifacts.

11.2 Evaluation of design science research

The criteria provided by Hevner et al (2004) for the evaluation of design science research are discussed in this chapter together with the author's self assessment of how the these criteria were met. The criteria for design science research evaluation are presented in section 11.2.1. Sections 11.2.2 through 11.2.6 provide a discussion of the author's self-assessment based on Hevner et al's (2004) assessment criteria.

11.2.1 Criteria for design science research evaluation

Manson (2006) has argued that criteria for the evaluation of research help researchers, reviewers, editors, and readers to understand the requirements for effective research. Hevner et al (2004) have provided seven guidelines for evaluating design science research as shown in table 11.1. Even though Hevner et al (2004) have advised against mandatory use of these guidelines, the author is of the opinion that in the absence of alternative guidelines at her disposal, these guidelines are suitable for stating the research contributions and conducting a self-assessment of the work done. The extent to which requirement number 2 (problem relevance) was met, was discussed in chapters 1, 2 and 3 of this thesis. The research that was conducted is assessed in the following sections.

Table 11.1: Criteria for the evaluation of design science research: adopted from Hevner et al (2004/5)

Criterion / Requirement	**Description**
1. Design of an artifact	Design science research for Information Systems must produce a useful artifact in the form of a construct, model, method or an instantiation
2. Problem relevance	The problem that the design science research is aimed at solving, must be technology-based, important, and relevant to some business function.
3. Design evaluation	The utility, quality, and effectiveness of the designed artifact must be rigorously demonstrated using well executed methods of evaluation.
4. Research contributions	Design science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies
5. Research rigor	The researcher must demonstrate that rigorous methods were applied in both the construction and evaluation of the designed artifact.
6. Design as a search process	The researcher must demonstrate that the available means were utilized well, in order to reach the desired ends while satisfying the laws in the problem environment (Satisficing).
7. Communication of the research	Design science research must be presented effectively to the intended audience.

11.2.2 Constructs, models and better theories

Requirement number 1 in table 11.1 refers to the design of artifacts. Constructs and models are two of the artifacts that design science research must produce (Hevner et al, 2004). The author claims in this thesis that one construct that arose from this research is the concept of decision rule-based search for feature subset selection. The author further claims that two constructs that arose out of this research are: pVn modeling, confusion graphs and the associated sparse confusion matrix property for aggregate modeling. Theoretical models are propositions expressing the relationships between the constructs / concepts of the research domain. Theoretical models were developed to express the relationships between the factors that affect the quality of selected features, and the factors that have an influence on the outcome of training dataset selection for aggregate modeling.

With reference to *better theories*, the experimental results were used to demonstrate the relationships between the various factors that affect predictive model performance. It will be necessary in future to conduct causation experiments (Cohen, 1995) to provide proof of these relationships.

11.2.3 Methods and instantiations

Methods and instantiations are two of the artifacts that design science research must produce (Hevner et al, 2004). A summary of the research contributions and proposed methods for feature selection, base model design, training dataset selection, and model aggregation for large datasets was given in sections 10.2 and 10.3. In this thesis, the author claims that the proposed methods and algorithms result in the selection of good feature ranking, good feature subset selection, design of highly competent base models, and selection of good training datasets for the base models. Furthermore, the proposed algorithm for model combination of 5NN base models (and KNN models in general) result in more effective resolution of conflicting predictions by the base models. Instantiation refers to the creation (implementation) of artifacts (constructs, models, methods) and demonstration that the artifacts can be implemented in a working system. All the methods and algorithms proposed in this thesis were implemented, tested and found to provide statistically significant improvements in predictive performance.

11.2.4 Rigorous design evaluation

Requirements number 3 and number 5 in table 11.1 are concerned with design evaluation, and the rigor that is applied to the evaluation processes. Hevner et al (2004) have stated that design evaluation involves the demonstration of utility, quality and effectiveness. Furthermore, the evaluation methods used to demonstrate (claim) the utility, quality and effectiveness of the methods and instantiations should also be evaluated. Hevner et al (2004) have further stated that designed artifacts should be evaluated using the methodologies that exist in the knowledge base for the field of research.

The evaluation methods that are available for predictive data mining originate from the area of Statistics, Machine Learning and Operations Research and were discussed in section 4.7. These evaluation methods enable the researcher to:

(1) Measure the predictive performance of a model in terms of overall predictive accuracy and error rate on all the classes, true positive, false positive, true negative, and false negative rates on each individual class.

(2) Measure the performance gains of using an aggregate model compared to using a single model.

(3) Conduct statistical tests, most commonly the Student's t-tests on means and F-tests on variance, to compare the predictive accuracy of two models.

(4) Conduct in-depth model analysis using ROC curves.

All the above methods were used for this thesis for the assessment of model performance. Machine learning research has traditionally concentrated on small datasets as exemplified by the datasets available from the UCI Machine Learning repository (Ascuncion & Newman, 2007; Blake & Merz, 1998). These datasets range in size from 100 instances to 10000 instances, and typically have a small number of predictive features. Researchers in machine learning have routinely used many small datasets (e.g. 30) to evaluate algorithm performance. However, as discussed in chapter 4, for experimental studies on aggregate modeling, bias and variance reduction, researchers have typically used small numbers of datasets ranging between two and nine datasets. The exception has been Ali and Pazzani (1996) who have used 30 small datasets. Performance evaluation using 30 small datasets can be feasibly conducted using a modest amount of time and computational resources.

Data mining poses new challenges in terms of evaluation. Typically, very large datasets are used as exemplified by the datasets available from the UCI KDD archive (Bay et al, 2000; Hettich & Bay, 1999). Datasets for data mining research range in size from 0.1 million instances to several million instances. Additionally these datasets have large numbers of potentially predictive features. In the author's opinion, the demonstration of rigor in evaluation, through the use of many very large datasets requires an excessively large amount of time and computational resources, which are not available to many researchers. In chapter 4 it was observed that experimental studies in dataset selection and aggregate modeling have been conducted by teams of researchers using between one and four very large datasets.

The author used a small number of datasets. Two small datasets (Abalone and Mushroom) and two large datasets (forest cover type and KDD Cup 1999) were used for feature selection. Two large datasets (forest cover type and KDD Cup 1999) and one small dataset (wine quality) were used for the training dataset selection and aggregate modeling studies. Twenty four models were created and tested for the three datasets, two algorithms, and four modeling methods. Many samples were taken from the large datasets, and the sample sizes used were larger than the typical dataset size for machine learning. Experiments were designed through the application of the scientific method and the evaluation methods listed above were employed.

11.2.5 Rigor and design as a search process

Requirements number 5 and number 6 in table 11.1 are concerned with the search process followed to arrive at good solutions for artifact design, and the rigor that is applied to the search process. Hevner et al (2004) have stated that rigor in the design process for design science research is derived from the effective use of the existing knowledge base (theoretical foundations and methodologies) of the field of research. A detailed assessment of the theoretical foundations of existing methods of dataset and feature selection was provided in chapters 2 and 3. A discussion was provided on how several existing theories can be applied to the task of designing feature selection and training dataset selection from large datasets for aggregate model implementation. The experiments presented in chapters 5, 7 and 8 were designed based on the assessments given in chapters 2 and 3 and the methodologies presented in chapter 4.

Hevner et al (2004) and Simon (1996) have observed that the design of artifacts is a search process aimed at the discovery of an effective solution to a problem. Hevner et al (2004) and Simon (1996) have characterized the design process as a generate-and-test cycle involving the generation of design alternatives and testing the alternatives against specific requirements. To the author's understanding, the generate-and-test cycle discussed by Hevner et al (2004) and Simon (1996) is identical to the scientific method that was discussed in chapter 4, and depicted in figure 4.3. Hevner et al (2004) have observed that an un-guided search for design alternatives would be intractable. It is usually prudent to employ heuristic strategies in order to generate designs for satisfactory solutions. In the field of Operations

Research, this approach is called *satisficing* (Simon, 1996). Heuristic search and *satisficing* for the scientific method are achieved through the cycle of: (*experiment-design)→(empirical-testing)→(empirical-observation)→(hypothesis-generation)→(experiment-design)*, as depicted in figure 4.3. The scientific method was followed for the studies reported in chapters 5, 7 and 8.

11.2.6 Research contributions for design science research

Requirement number 4 in table 11.1 is concerned with research contributions. Hevner et al (2004) have observed that any assessment of a research activity must answer the question: *'What are the new and interesting contributions?'* Hevner et al (2004) have stated that design science research must provide one or more of the following contributions: *design artifact, foundations* and *methodologies*. For this thesis, the author claims that the design artifacts that were discussed in sections 11.2.2 and 11.2.3 are research contributions to the field of predictive data mining. *Foundations* refer to the knowledge base of the field. The author further claims that the algorithms presented in chapters 5, 6 and 8 for feature selection and aggregate modelling are contributions to the field of predictive data mining.

Table 11.2 provides a summary of the new algorithms proposed in this thesis. The guidelines for feature selection and training dataset selection, new modeling methods, and theoretical models discussed in chapter 10 are a research contribution to the field of predictive data mining.

Table 11.2: Summary of new algorithms

Algorithm category	Location	Description
Feature selection	Fig. 5.3	Decision rule-base search algorithm for heuristic search of the best feature subset
OVA modeling	Fig. 6.3	Algorithm for combining base model predictions for the See5 algorithm and for classification trees in general
pVn modeling	Fig. 6.4	Algorithm for combining base model predictions for the 5NN algorithm and for the KNN algorithm in general
	Fig. 8.3	Algorithm for class selection of pVn base model
	Fig. 8.6	Modified algorithm for class selection for pVn base models

Methodologies refer to the creative development and use of new evaluation methods and evaluation metrics. A modified version of Ali and Pazzani's (1996) performance improvement measures were presented in chapter 6 and used extensively for chapters 7 and 8. A modified version of Provost and Domingo's (2001) VUS estimate

was presented in section 9.2. In this thesis the author claims that these modified measures provide a modest extension to existing evaluation metrics for predictive modeling.

11.3 Limitations of the proposed dataset selection methods

Toulmin's argumentation model (Toulmin el al, 1979; Toulmin, 1958) which explains the structure of claims in scientific discourse, and Ngwenyama's (2007) analysis of scientific research claims were introduced in chapter 1. Recall that claims are supported by *data* (evidence*), warrants* (rules of inference) and *backing* (authoritative sources for *warrants*). Two additional components in Toulmin's model are *qualifiers* and *rebuttals*. *Qualifiers* are used to limit the strength of a *claim* and *rebuttals* provide an elaboration for the *qualifiers*. The claims made in this thesis concern the effectiveness of feature selection and training dataset selection, and aggregate modeling methods as discussed in chapters 5 to 8 and summarised in chapter 10.

A claim was made in chapter 5 that the use of many samples to measure class-feature and feature-feature correlations is an effective method for the accurate measurement of these correlations. However, the datasets used in the studies were of moderately high dimensionality. In practice there are many problem domains for which the dimensionality of the datasets are extremely high. The use of many samples to measure correlations coupled with robust correlation measures with quadratic time complexity is a daunting task. This issue was not addressed in this thesis and is left for future work. The *qualifier* for the *claim* is that the proposed methods of using many samples to measure correlations *are only appropriate when* the dimensionality of a dataset is not very high.

Claims were made in chapters 7, 8 and 9 that the proposed methods of base model design and training dataset selection for OVA and pVn modeling result in aggregate models that have a higher level of predictive performance compared to single *k*-class models. A *qualifier* was stated in section 8.5.3 that the proposed methods for boosted OVA and pVn model design are *only appropriate when* a dataset has a single *k*-class model confusion matrix with the sparse confusion matrix property. The situations where a non-sparse confusion matrix can be transformed into a sparse confusion matrix were also given in section 8.5.3.

11.4 Chapter Summary

The research outputs and claims of contributions in this thesis were assessed in the context of design science outputs and research contributions. The limitations of the proposed methods were also discussed. The conclusions for the thesis are presented in the next chapter.

Chapter 12
Conclusions

> 'You are my life. In you my peace, in you my joy, in you my strength, in you my God.'
> (Benjamin Dube, 2007)

12.1 Summary of the thesis

The central argument of this thesis is that, it is possible for predictive data mining to systematically select many dataset samples and employ different approaches (different from current practice) to feature selection, training dataset selection, and model construction. When a large amount of information in the large dataset is utilised in the modeling process, the resulting models should have a high level of predictive performance and should be reliable.

The discussions of chapters 2 argued that there is a need for methods for training dataset selection from large datasets, using as much data as possible with the objective of reducing the bias and variance components of the prediction error. The discussions of chapter 3 argued for the need for feature selection from large datasets, with the objective of using as much data as possible in order to reliably measure the class-feature and feature-feature correlations used in the feature selection process.

The experimental results of chapter 5 demonstrated that the use of the mean values for the correlations, obtained through the use of many samples, robust measure of correlations, and validation methods such as the use of fake variables, results in the identification of features which are relevant for the prediction task. The experimental results of chapter 5 also revealed that the incorporation of domain-specific definitions of the meaning of *low, medium* and *high* correlation into a feature subset search procedure results in the selection of good feature subsets for the prediction task at hand. The experimental results of chapters 7, 8 and 9 demonstrated that the use of the proposed methods for base model design and training dataset selection for OVA and pVn aggregate modeling has the potential to produce models which have a higher level of predictive performance compared to single models.

12.2 Conclusions and reflection

From a computational perspective it can be argued that the methods proposed in this thesis provide the following desirable outcomes: Firstly, the methods result in the use of large amounts of data which provide a large amount of information to the modeling process. Secondly, the methods for OVA and pVn modeling lead to the avoidance of un-necessary computations since the modeling effort is aimed at the creation of models that have a potential to increase predictive performance. From a statistical perspective it can be argued that the proposed methods provide the following desirable outcome: The methods result in the use of large amounts of data and at the same time avoid the problems of overfitting, data dredging and the modeling of phantom (chance) structure. From an Operations Research perspective it can be argued that the proposed methods provide the following desirable outcome: One of the uses of ROC analysis is used to determine the optimal operating point for a predictive model. The proposed OVA and pVn modeling methods have the potential to produce predictive models with higher optimal performance compared to single models.

It has been demonstrated that the use of large amounts of data with the methods proposed in this thesis, has the potential to provide predictive models with a high level of predictive performance. In general, no single method can be claimed to be suitable for all datasets and for all algorithms. Schaffer (1994) has argued that no single strategy for machine learning is better at generalisation (prediction) than all other strategies for all problem domains. In his study of noise-free datasets, Wolpert (1996) has demonstrated through the *no free lunch theorems for machine learning* that all algorithms are equivalent on average, in terms of predictive performance. The foregoing arguments can be easily extended to other computational domains. With the foregoing observations in mind, the author does not claim that the proposed methods will provide effective solutions for all data mining application domains. In order to establish the extent of applicability for the proposed methods additional empirical studies as discussed in the next section, will have to be conducted in future.

12.3 Future work

It was observed in chapter 5 that predictive features can be eliminated when robust correlation measures are used even when such features are good predictors for one or more local areas of the instance space. It will be useful in future to conduct studies for the identification of locally predictive features which are predictive of real structure as opposed to phantom (chance) structure. It was also observed in chapter 5 that predictive features for severely under-represented classes may be eliminated when robust correlation measures are employed. In future it will be useful to study feature selection methods that directly address this problem.

It was observed in chapters 5 and 10 that use of many samples to measure correlations coupled with robust correlation measures with quadratic time complexity is not a feasible approach for the estimation and validation of class-feature and feature-feature correlation coefficients for datasets of very high dimensionality. It will be useful in future to study feasible and reliable methods of correlation measurement for datasets of very high dimensionality.

The confusion matrix was used for the experiments of chapters 7 and 8 as a basis for the identification of confusion regions for a classification task. It will be useful in future to investigate other methods for the identification of confusion regions. Confusion graphs were used as input to the algorithms for determining the design of pVn models. The weights for the arcs of the confusion graphs were not used in the algorithm's decisions except in the case where a maximally connected graph had to be pre-processed. It will be useful to investigate how the arc weights in a confusion graph can be used to fine tune the decisions of these algorithms.

The dataset selection and aggregate modeling methods proposed in this thesis were directed at multi-class problems, and are not directly applicable to 2-class prediction problems unless a dataset is pre-processed through cluster analysis as discussed in section 8.5.3. It will be useful in future to investigate how the proposed OVA and pVn base model design and training dataset selection methods could be extended to 2-class problems.

It was stated in sections 8.5.3 that the proposed base model design and training dataset selection methods for boosted OVA and pVn aggregate models are only

applicable when the single *k*-class confusion matrix for a dataset has the *sparse confusion matrix* property. It will be useful to investigate different problem decomposition methods (different from OVA and pVn) for such datasets. For such problem decomposition methods it will be necessary to design training dataset selection methods for bias error reduction.

It was observed in chapters 7 and 10 that if an algorithm for the combination of base model predictions is able to resolve conflicting (tied) predictions then a high level of predictive performance is realised for an aggregate model. This was shown to be the case for 5NN classification. It will be useful in future to investigate methods of resolving conflicting (tied) predictions for classification tree algorithms.