Data Mining in the Real World: Experiences, Challenges, and Recommendations

Gary M. Weiss
Department of Computer and Information Science, Fordham University, Bronx, NY, USA

Abstract - *Data mining is used regularly in a variety of industries and is continuing to gain in both popularity and acceptance. However, applying data mining methods to complex real-world tasks is far from straightforward and many pitfalls face data mining practitioners. However, most research in the field tends to focus on the algorithmic issues that arise in data mining and ignores the human element and process issues that are often the cause of these pitfalls. While there are some papers on data mining experiences and lessons learned, they are quite rare, especially in the research community. The purpose of this paper is to begin to fill in the "gap" between data mining methods and the* practice *of data mining. This paper describes many of the experiences of the author as a data mining practitioner, highlights the issues that he encountered while in industry, and provides a number of strategies and recommendations for dealing with these issues. This paper should benefit both practitioners and researchers but hopefully, more than anything, it will foster discussion on the practical data mining issues that are all too often ignored.*

Keywords: data mining applications, data mining lessons

1 Introduction

Data mining is fundamentally an applied field, driven more by a class of problems (e.g., classification, clustering, etc.) than by a specific set of methods. Nonetheless, most published work in the field focuses almost exclusively on data mining methods and algorithms. Although an increasing amount of work is beginning to focus on the characteristics of real-world problems that makes data mining difficult (class imbalance, unequal misclassification costs, etc.) there is still relatively little work that describes the many practical issues that arise when addressing real data mining problems. This paper takes one small step toward remedying this deficiency by discussing issues that the author has encountered during his years as a data mining practitioner. The hope is that a discussion of these issues will be useful to both data mining practitioners and researchers. In addition to raising a variety of issues, the paper also provides a number of strategies and recommendations. However, it should be noted that this paper focuses on the author's personal experiences, which certainly may differ from the experiences of other practitioners. Furthermore, this paper also tends to focus more on organizational and process issues than on technical issues.

As just mentioned, this paper is not meant to be a comprehensive accounting of all of the issues that may arise when tackling difficult, real-world, data mining problems. Rather, it focuses on the author's experiences as a data mining practitioner. Because of this, the author's work history is relevant and is briefly described here. Prior to moving into academia the author worked for over eighteen years at AT&T Bell Labs and AT&T Labs, of which the last nine years were spent either doing data mining or in activities related to data mining. In one position the author designed and implemented an expert system to maintain central off switching equipment [15] and was able to supplement the knowledge acquired from domain experts with knowledge extracted mined directly from telecommunication alarm data [14]. In his next position the author served as a technology expert, promoting and educating others on advanced technologies, including data mining technology. In his last position at AT&T, the authors spent five years in a market analysis group applying data mining methods to solve business and marketing problems. The author has also gained some industry experience since joining Fordham University, via a grant from a large multinational financial institution to investigate the use of data mining in that domain. Useful insights have also been gained from graduate students who work in industry as professional data analysts.

This paper shares the author's experiences and describes many of the issues that he has faced. These issues and experiences are organized around several topics. Organizational issues that impact data mining, including various "political" issues, are described in Section 2. There are many issues involved with obtaining suitable data and Section 3 covers this issue. Section 4 describes general issues related to the process of data mining, including the selection of data mining tools. Finally, Section 5 summarizes the issues and experiences described throughout this paper as well strategies and recommendations for dealing with these issues.

2 Organizational Issues

There are many organizational and non-technical issues—such as political issues—that may arise in a data mining project. The first thing that a data mining practitioner needs to be aware of is the organizational context in which he operates. Are the data mining practitioners an integral part of the organization that is initiating the project, are they in another part of the same company (acting as internal consultants), or are

they acting as outside consultants? Typically the closer the data mining practitioners are to the initiators of the data mining effort, the greater the support they can expect—but that is not always the case and there can certainly be resistance to data mining even within one's own organization. Thus, what matters is not just the organizational context, but the support that the data mining practitioners receive from the organization sponsoring the work. As one group of data mining practitioners explains [2] "never underestimate the power of politics and turf battles," a sentiment this author certainly agrees with.

The author has encountered a variety of organizational issues during his tenure in industry. However, it is also important to understand the underlying *causes* for these issues, so that one can predict when these issues are likely to arise and possibly develop strategies for addressing or avoiding them. A key organizational issue relates to obtaining the data necessary for data mining (Section 3 covers the non-organizational issues related to data acquisition). Because the meaning of the data fields is often not fully documented or is only understandable to domain experts, it is usually critical to have access to these domain experts. Data mining expertise simply cannot make up for a lack of domain knowledge—something the author has found out the hard way. Furthermore, getting sufficient access to the domain experts is often even more difficult than getting access to the data—even when the domain expertise resides in the same company. There are a number of reasons for this—all of which in the author's case were probably exacerbated by the fact that his company was in a consistent "downsizing" mode during his years as a data mining practitioner. These reasons include:

- Job security: *sharing* data and expertise may make you or your organization less critical to the business.
- Power: as the adage goes, "knowledge is power" and the more knowledge you control the more powerful you are. Most groups within a business want to retain or increase their power and not dilute it by sharing critical information. Organizations also tend to want to keep work for themselves and increase their "head count" (i.e., number of employees).
- Limited Time: Time is a valuable resource and sharing knowledge really does require significant time. Also, the most knowledgeable domain experts—which are the ones you really need access to—are typically the most valued and in demand.
- Lack of Budget: For many data mining projects there is either no budget or a very small budget to reimburse the domain experts for their time. Thus, the organization with the necessary information may not be fairly compensated for access to their domain experts.
- Pride: A group often believes that they are doing the best job possible and that there is no benefit in bringing in outsiders, potentially with "new-fangled" and (in their eyes) unproven methods.

While some of these reasons may not be consistent with the best interests of the business, they are rational from the perspective of the individual workers and should not be underestimated or ignored. In the author's experience these issues arise in virtually every project. These issues are certainly not specific to data mining and in fact the author has encountered these same issues when working on expert systems [15], which, like data mining, requires a lot of domain expertise and can lead to a loss of control of critical knowledge and lost jobs within an organization. As a concrete example, in one project the author needed to mine the call detail data that describes every individual phone call that traverses the telecommunications network, but many of the fields were hard to interpret and poorly documented, so additional domain expertise was required. The person with that domain expertise was located in another organization within the same company and had previously functioned as a consultant to the author's organization—and had received funding for that. Given that "downsizing" was occurring, the domain expert was not motivated to share his domain expertise, acquired painstakingly over many years, and thus it was difficult to obtain the necessary information.

The organizational issues just mentioned may not have easy solutions, but there are certain actions that can be taken. First, assess the amount of organizational support before starting the project and, if at all possible, get formal commitments for the support before starting the project. Also, make sure that the necessary domain experts are allocated to the project and funded at the appropriate level, and, if feasible, transferred into the organization, even if only on a temporary basis.

There are a variety of other organizational and political issues that can also arise. One issue relates to a bias that is sometimes found against data mining. This bias is probably due to our field being relatively new and the unjustified hype that occurred as data mining began to become popular. While the author found relatively little resistance to data mining while at his former employer, many of the graduate students from the Economics department, who take his data mining course, have specifically noted this bias when working in industry. This is undoubtedly due to the fact that Economics has traditionally been dominated by Statistics and the influence of Computer Science has been relatively minimal, at least until recently. Some of these students have received very negative comments about data mining, especially as they try to utilize their new skills at work, including the fact that "it just does not work." The best way to combat the bias against data mining is through the deployment of successful applications and education. It certainly is a good idea for data mining practitioners to be aware of the commonalities and differences between data mining and statistics and this can be done by reading articles on this topic by Friedman [5] and Hand [6].

Given these and other issues, for data mining to succeed it often needs to yield significant, non-incremental, improvements. Otherwise there may be too much inertia in the organization to deploy new processes. The author has found this to be true, especially when promoting other "new" technologies

such as expert systems and object-oriented programming and design. Many projects the author has been involved in were never completed because the benefits were not clear enough to warrant continued work. Other data mining practitioners have also found this to be true [2], as evidenced by the statement "The data mining algorithm has to perform demonstrably and considerably better than an existing system." Thus it is important to assess the potential upside of a data mining project early on and determine whether the potential benefit is sufficient to ensure adoption of the project. It is also important to balance short-term and long-term needs since short-term benefits can help maintain support for a project. This should be kept in mind as the work in being planned and because of this a phased approach is often best. This approach is advocated by one group of data mining practitioners [8] when they recommend that one should "crawl, walk, run." As they point out and as the author noted for himself, in many cases the customer may not even require true data mining but instead just needs some fancy reports. In this case it is generally best to satisfy their immediate needs and then move on to a more complex analysis of the data, if necessary.

Many projects, however, are dropped not due to inertia, but due to an even more significant issue—the project was never *bought into* by the people in charge. This might sound silly in a business environment, but it occurs all of the time. It is especially prevalent when dealing with new technologies and in companies that are technology driven, where the technology experts (as in the author's case) often reside in separate organizations, or *laboratories*. In this situation it is often not the case that the business customer recognizes a need and seeks expert advice. Rather, a more typical scenario is that the technology expert actively tries to find a customer and business problem that matches his expertise.

The practice just described may seem like a poor business model—and perhaps it is—but there is some justification for it when the customer is not knowledgeable about the new technology and thus may not even recognize suitable business problems. This is especially true for data mining. In organizations where this practice exists the data mining practitioner may need to spend a significant amount of time educating the customer, trying to elicit good business problems, and selling his ideas. Thus, a data mining practitioner should have good communication skills and also be a good educator. Unfortunately it is often difficult to get "buy in" for a project early on, and thus one has a tendency to just jump into the project, without a firm commitment from the customer. The hope is that one can achieve good results quickly and use this to then get the necessary commitment. This often leads to a lack of resources (such as time and expertise from the customer) which may ultimately harm the project. But there are benefits to this model for data mining practitioners and their organizations. In the author's organization if you achieved good results and managed to get the customer to buy into a project, then you could repeat the analysis periodically and receive substantial funding with relatively little effort. This funding could then be used in turn be used to subsidize the exploratory phase of other data mining projects.

There are many other organizational issues that could be discussed. One issue involves who should perform the data mining analysis—should it be done "in house" or by an external consultant. At the start of the authors career it was perceived that his employer had the "not invented here syndrome", where everything that could be done internally was done internally (e.g., one organization even developed its own operating system). However, as competitive pressures mounted and corporate cultures changed, the company and most other companies moved away from that philosophy. Now the "don't reinvent the wheel" philosophy seems to dominate, where any work that can be outsourced is outsourced. This certainly includes data mining, since many companies cannot support their own data mining organizations or experts. One should be aware of the philosophy of their organization and its implications. The author recently acted as a consultant for a large multinational bank for the initial stages of a project, which could have been done internally, but the decision was made to outsource it all—at great expense. While there are advantages to outsourcing data mining work, there are many disadvantages, some of which have already been mentioned (e.g., the larger the distance to the customer the more resistance). However, one issue that should not be forgotten is that domain expertise is very important and difficult to acquire. Thus, if the knowledge can be kept internally, maintained, and reused, it may be less costly in the long run.

The author's industry experience involved working with data analysts from very different background. Most were statisticians or, more specifically, econometricians. Interacting with people who have diverse backgrounds is very useful and provides a good environment for learning. This is especially useful when some problems are more suited to one discipline over another. For example, much of the data that the author has mined in industry is time series data and very often statistical techniques are more appropriate for handling this type of data. Thus, it can be beneficial for organizations that are involved with data analysis to employ people with diverse backgrounds, rather than isolating them in different groups—which is a common practice.

3 Obtaining and Preparing Data

Data acquisition is one of the key stages in the data mining process [4] and in the author's experience this is one of the most problematic stages. Section 2 focused on some of the organizational and political issues that may complicate data acquisition but in this section we focus on some additional issues that complicate this step. To appreciate the importance of the data acquisition phase it is important to understand that data acquisition is typically the most time-consuming part of the data mining process. The author's experiences—which are generally consistent with that of other data mining practitioners—indicate that significantly more than half of the overall data mining effort is spent obtaining, preparing, preprocessing, and transforming the data. It is critical that new data mining projects budget for this time and provide the

resources necessary to acquire the data. For those who are interested in learning more about data preparation, an entire text is dedicated to this topic [9].

The data that one would like to have for data mining often is not available. In some cases it may be possible to modify business processes to acquire the data, but often that is not practical. For example, when analyzing telecommunication network data the author sometimes found that useful information was not stored, but since modifying the software on telecommunication switches is both costly and risky, it was not even feasible to start collecting the desired data on an ongoing basis. The best practice is to carefully consider what data may potentially be useful when new data sources becomes available and store as much as possible, or, as Kohavi et. al [3] suggest, "Collect the right data, up front." Identifying the "right data" is not easy since it is difficult to envision all potential uses for data, but corporations are finally taking the time to do this because they now recognize the business value of data.

Having data stored is important, but it is equally important—if not more important—to make sure that the data is readily *accessible*, given the many organizational issues mentioned in Section 2. There has been a shift toward easier access to data in industry over the past decade, as many corporations have moved from hundreds of separate, independent databases to a few large, carefully maintained, data warehouses with simple, uniform, (often web-based) user interfaces. For large corporations with many legacy systems this was extremely costly to implement, but ultimately is well worth the effort. Unfortunately, much of the data stored in typical data warehouses is summarized data, whereas in data mining one typically requires access to the lowest level, non-aggregated, data. Until recently it was difficult or impossible to permanently store these huge amounts of data (e.g., the AT&T call record database contains over 1.9 *trillion* records) and as a consequence one would only have access to the full data for a limited time period, which can be problematic if one is trying to track changes over time or needs to model phenomena that occur only occasionally (e.g., phone traffic on Mother's day). However, over the past few years it has finally become feasible to store hundreds of Terabytes of data, so now even non-aggregated data can often be kept in its entirety. Having this data, clearly documented and easily accessible, can dramatically reduce the time required during the data acquisition phase of data mining.

One problem that the author has encountered repeatedly, which is often overlooked, relates to combining data from multiple sources. This combination requires a common key, but such a key is not always available. In particular, in a large company organized into many independent business units (some of which may have formerly been independent companies), different databases often have different ways of identifying customers. Thus even the most basic merging of data can be problematic. For example, a customer like IBM might have many variations of its name and the contact information, which is often used as a secondary key to uniquely identify a customer, may vary. This "name matching" problem was also encountered by the author when working with a large financial institution. In each instance where this problem was encountered by the author, the problem required months or years of effort to address and the development of specialized name-matching software. This problem is significant and can sometimes be avoided by developing consistent terminology within a company and carefully reviewing database schemas, but even this is not sufficient to avoid all problems due to unanticipated changes such as mergers and acquisitions.

One of the key steps in the data mining process [4] involves preparing the data so that it is suitable for data mining. Very often the raw data must first be transformed before it is useful and this was especially true in the author's experience, largely due to the fact that much of the data he mined was temporal in nature (i.e., data streams) and could not be fed directly into conventional data mining prediction algorithms (e.g., decision trees, rule-based). For instance, call detail records are essentially generated in real-time by a telecommunication network as a data stream, but in order to mine information about individual customers (represented by phone numbers) all of the calls associated with a phone number must be aggregated into a single record. This is typically done by introducing a host of summary features, such as *average call duration* and *average number of calls per day*. The aggregation process must be done carefully to ensure that the information most critical for data mining is not lost and this is where domain expertise and insight can be critical—and where creativity may enter the process via the construction of new features. As an example, one data mining problem that involved trying to identify telemarketers based on their calling behavior. To help address this problem we introduced a feature that measured the *dispersion* of a user's phone calls with respect to different area codes, since most traditional residential and business customers call only a relatively few area codes. This feature turned out to be very useful in building the predictive model. In general, feature construction or feature engineering is a critical step which can greatly impact the success or failure of a data mining project. The issues with data streams are also not unique to the telecommunications industry and arise in many different domains, such as business (daily stock prices) and medicine (cardiac monitoring).

Data quality is another key issue and it is critical that a data mining practitioner validate the data and clean it as necessary. The author has often found errors in the data and has also found that the documentation that describes the data is often incorrect or out of date. Examining the distribution of each data field is critical and can quickly identify errors—as well as outliers. As an example, one project the author participated in involved analyzing the behavior of "800" number calls, which incur no cost to the call originator. After sorting these 800 numbers by total number of calls per month, it turned out that a few of these numbers accounted for a large percentage of the total calls (a few of the 800 numbers were associated with millions of calls per month). As it turned out, these numbers were calling card numbers owned by the telecommunication company and anyone who used one of its calling cards had to call this number. We then removed these

numbers from the analysis since we were only interested in predicting the phone usage of 800 numbers owned by "true" customers. The key lesson is to understand your data and ultimately the best way to do this is to spend time looking at the data records and the distribution of values for each field.

Another data-related issue that the author has repeatedly encountered concerns class imbalance [13]. Many prediction problems, including the prediction of telecommunication equipment failures [14], are extremely rare. The issue of dealing with "unbalanced data" is not at all uncommon in data mining, but yet standard data mining algorithms, which are geared toward maximizing predictive accuracy, usually perform poorly in these cases. Partially for this reason the author developed a custom tool, Timeweaver [12], which was designed to perform well for unbalanced classification problems, but due to the inherent complexity of this problem the author was forced to tackle the easier problem of predicting circuit pack failures rather than the much rarer catastrophic failures of entire switches. Fortunately there has been a great deal of work in recent years in mining unbalanced data, including two workshops on this topic [1] [7]. One thing that is important in these cases is to try to obtain accurate cost information, so that cost-sensitive learning can be used to compensate for the class imbalance. Since in this case the cost of a false negative will almost always be higher than the cost of a false positive (where the rare class is the positive class), cost sensitive learning will tend to cause more of the rare cases to be predicted, which is usually desirable. Unfortunately, it is often difficult to obtain accurate cost information, which is why this information is usually not provided. Nonetheless, every attempt should be made to obtain this cost information, even if one is only able to obtain rough estimates for these costs.

The acquisition of data from external sources (e.g., census data, data about business from D&B, etc.) represents an opportunity that is often overlooked. In many cases the data may not directly address the underlying data mining problem in an obvious way, but may nonetheless lead to significant improvements. This also provides an opportunity for the practitioner to be creative. As one example, many of the author's analyses have involved businesses at specific locations. Based on data taken from the U.S. Census, the author has been able to supplement this data with aggregate data associated with the businesses' location. For example, depending on the specific data mining problem, it might be useful to know if the geographic area has a higher than normal percentage of high-tech companies or people with graduate degrees. Practitioners should thus be open to the use of all sources of data, many of which can be acquired without cost.

Data, however, is often acquired only with a cost, even if that cost is not in the form of payments to a third party. As discussed earlier, it may take time to acquire the data, clean the data, and transform the data. Thus, one decision that the author encountered regularly was how much data to acquire, clean, and process. This topic is not something that is often discussed in the research literature, although it has been studied [16]. While one could generate learning curves by sampling different training set sizes from the currently available training data, a simpler but effective strategy is to just generate a few points on the learning curve below but close to the current training set size, in order to estimate the benefit of obtaining additional training data. Then, depending on the slope of the learning curve near the current training set size and the cost of acquiring additional data, one can determine whether further data acquisition is warranted.

4 Data Mining Process

This section describes issues related to the general process of data mining and the selection of tools for performing data mining. The data mining process is an iterative process and experience has shown that it is important to take this into account. Very often data mining will not go as expected and one must respond based on the feedback gained along he way. One strategy is to build a quick prototype, even though this generally requires that a variety of data quality issues must (initially) be ignored. Generating *some* results quickly, however, is very useful since this will ensure that you have a well defined problem and well thought out evaluation criteria. The reasons for building quick prototypes in data mining are largely the same ones for building fast prototypes for software systems, where the "waterfall" model of software development often fails because it does not easily accommodate feedback and changing requirements. Kohavi et al. [8] essentially advocate this strategy when they advise "build simple models first."

Another reason for building quick prototypes is to establish a baseline against which to measure future improvements. If complex modeling or highly engineered features fail to yield improvements over simple methods and features, then the complex ones should not be employed. It is also important to ask if the initial model—which may not even have been generated using data mining—is sufficient for solving the problem. Past experience has demonstrated that quick prototypes sometimes are satisfactory. In one case the business problem was to determine the fraction of phone lines and phone minutes associated with voice, data, and fax. The initial plan was to build a training set for various usage segments (where the segments were based on the number of monthly phone minutes), train a predictive model for each segment, and then apply the predictive model to the "universe" of phone lines. As it turned out, the segments with very large phone minutes were small enough to be sampled completely and our automated tool, which automatically dialed the lines to determine the type of phone usage, could, in the time allotted, label 10,000 lines from each of the segments with lower phone usage. Given all of this data and our random sample for each segment, no predictive model was required. Instead, we just scaled up the numbers to account for the known number of lines in each usage segment. In other cases experience has shown that only very simple predictive models, sometimes using only a few variables, are necessary.

There are many different data mining tools that are available and care should be taken when choosing a tool to use.

Because the author has not performed an extensive evaluation of these tools, comprehensive guidance as to which tool is best can not be provided. However, the author has used a number of different tools and while in industry used both SAS Enterprise Miner [10] and SPSS Clementine [11], two of the more powerful and widely used commercial data mining packages. These tools are comprehensive data mining suites that support a large number of data mining methods. Based on experience it is preferable to use suites such as these rather than stand-alone data mining tools that only support a single method. The advantage of using one of these (or similar) tools is that they allow one to easily generate multiple models, compare the performance of different models, and combine multiple models via an ensemble. These packages also provide basic support for data exploration and visualization, as well as methods for data transformation. In short, these tools provide at least some support for all parts of the data mining process [4]. While these commercial suites can be expensive, free alternatives do exist, such as WEKA [17], an open source data mining package from the University of Waikato that has many of the features of Clementine and Enterprise Miner. Others, including some of the author's students, have also had good experiences with R (http://www.r-project.org/), an open source statistical programming environment that contains a large number of data mining methods.

5 Summary and Recommendations

This paper discussed the personal experiences of the author while acting as a practitioner of data mining in an industrial setting. A number of issues were raised—issues that often are not discussed in research papers on data mining. Hopefully this discussion will provide some insight into the challenges one may encounter when using data mining to solve complex real-world problems. This paper also outlined strategies for addressing the issues and challenges that were raised and these challenges and recommendations are summarized in Appendix 1, provided at the end of this paper after the references. Hopefully this paper will stimulate further discussion, and research, on the *practice* of data mining.

6 References

[1] N. Chawla, N. Japkowicz and A. Kolcz, (editors), in *Proceedings of the ICML Workshop on Learning from Imbalanced Data Sets*, 2003 [online]. Available: http://www.site.uottawa.ca/~nat/Workshop2003/workshop2003.html.

[2] T. Dasu, E. Koutsofios, and J. Wright, "Zen and the art of data mining," in *Proc. of the KDD 2006 Workshop on Data Mining for Business Applications*, 2006, pp. 37-43.

[3] T. Dasu and G. M. Weiss, "Mining data streams," in *Encyclopedia of Data Warehousing and Mining, second edition*, J. Wang Ed. Kluwer Academic Publishers, 2008, pp. 1248-1256.

[4] U. Fayyad, G. Piatesky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp.37-54, Fall 1996.

[5] J. Friedman, "Data mining and statistics: what's the connection," *29th Symposium on the Interface: Mining and Modeling Massive Data Sets in Science, Engineering, and Business*, 1997.

[6] D. Hand, "Data mining: statistics and more?" *American Statistician*, vol. 52, no, 2, pp. 112-118, May 1998.

[7] N. Japkowicz (editor), *Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets* , AAAI Tech Report WS-00-05, July 2000.

[8] R. Kohavi, L. Mason, R. Parekh, and Z. Zheng, "Lessons and challenges from mining retail E-commerce data," *Machine Learning*, vol. 2, no. 1-2, pp. 83-113, Oct.-Nov. 2004.

[9] D. Pyle, *Data Preparation for Data Mining*. San Diego: Morgan Kaufmann, 1999.

[10] SAS Corporation, "SAS Enterprise Miner." [http://www.sas.com/technologies/analytics/datamining/miner]

[11] SPSS Corporation, "Clementine." [http://www.spss.com/clementine].

[12] G. M. Weiss, "Timeweaver: a genetic algorithm for identifying predictive patterns in sequences of events", in *Proc. of the Genetic and Evolutionary Computation Conference*, 1998, pp. 718-725.

[13] G. M. Weiss, "Mining with rarity: a unifying framework," *SIGKDD Explorations*, vol. 6, no. 1, pp. 7-19, June 2004.

[14] G. M. Weiss and H. Hirsh, "Learning to predicting rare events in event sequences," in *Proc. of the 4th International Conference on Know. Disc. And Data Mining*, 1998, pp. 359-363.

[15] G. M. Weiss, J. P. Ros, and A. Singhal, "ANSWER: Network Monitoring Using Object-Oriented Rules," in *Proc. of the 10th Conference on Innovative Applications of Artificial Intelligence*, 1998, pp. 1087-1093.

[16] G. M. Weiss and Y. Tian, "Maximizing classifier utility when there are data acquisition and modeling costs," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, pp. 253-282, Oct. 2008.

[17] I. H. Witten and E. Frank, *Data Mining: practical machine learning tools and techniques, 2nd Edition*, Morgan Kaufmann, San Francisco, 2005.

Appendix 1: Summary of Issues and Recommendations

The Table below lists the main issues discussed in this paper and then provides a summary of the recommendations associated with each of these issues.

Issue: lack of access to data and domain expertise (due to concerns with job security, pride, power and limited resources) • Provide sufficient budget and resources for acquiring the data and include time with domain experts. • If possible, have domain experts assigned to your organization, even if temporarily.
Issue: bias/fear of data mining • Education and understanding. Avoid hype. Read Friedman [5] and Hand [6].
Issue: organizations have inertia and are often reluctant to adopt new methods and technologies • Evaluate projects carefully and focus on ones that may yield clear and substantial improvements. • Used a phased approach that will provide some short-term "wins."
Issue: may often be no up-front "buy-in" from those who are responsible for funding the project. • Data mining practitioners should be a good educator, communicator, and salesman.
Issue: data may have errors and/or documentation can be out-of-date or just incorrect • Examine data distribution for each field and determine whether values make sense; investigate a sample of the outliers. • Select some examples at random and verify that the data values make sense.
Issue: it is often it is difficult to merge information from multiple data sources due to lack of unique key • Try to adopt company-wide standards (this is often not sufficient due to mergers and acquisitions). • Develop tools to perform the matching process (unfortunately this can be expensive and time consuming).
Issue: data may not be represented at the level needed for data mining • Transform the data to the appropriate level. If aggregation is necessary, preserve the most critical information. • Be creative and use domain knowledge and/or exploratory data analysis to come up with useful features.
Issue: class distribution of data is often highly unbalanced • Use appropriate evaluation metrics (avoid accuracy). Perhaps use ROC analysis if cost information is not available. • Try to acquire accurate cost information or at least reasonable estimates and then employ cost-sensitive learning.
Issue: good training data is costly to obtain • Generate learning curves to determine whether it may be profitable to obtain additional training examples. • Try to incorporate secondary data sources (some of which may be free), such as U.S. Census data.
Issue: the data mining problem and/or the evaluation criteria may not be well defined. • Build a quick prototype in order to get quick feedback. Customers respond best given something concrete to evaluate.
Issue: what data mining methods and tools should be used? • Data mining suites that support multiple data mining methods and the full data mining process are generally best. • No one data mining method is best, so try a variety of methods. Prefer simple methods.