FOCUSING ON THE DATA IN DATA MINING: LESSONS FROM RECENT EXPERIENCE

Neal J. Rothleder, Earl Harris, Eric Bloedorn

The MITRE Corporation
{neal, esharris, bloedorn}@mitre.org

The use of data mining is growing rapidly. The number of data mining consultants, as well as the number of commercial tools available to the "non-expert" user, are also quickly increasing. It is becoming easier than ever to collect datasets and apply data mining tools to them. As more and more non-experts seek to exploit this technology to help with their business, it becomes increasingly important that they understand the underlying assumptions and biases of these tools. There are a number of factors to consider before applying data mining to a database. In particular, there are important issues regarding the data which should be examined before proceeding with the data mining process. While these issues may be well-known to the data mining expert, the non-expert is often unaware of their importance. In this paper, we will focus on three specific issues, and illustrate each through the use of examples taken from our recent experiences. For each issue, we provide insight into how it might be problematic and suggest techniques for approaching such situations.

1. INTRODUCTION

The purpose of this paper is to help the non-expert in data mining better understand some of the important issues of the field. We are particularly concerned with characteristics of the data which may affect the overall usefulness of the mining results. Some recent experiences, and the lessons learned from them, are described. These lessons, together with the accompanying discussion, will help to both guide the data collection process and better understand what kinds of results to expect.

The use of data mining is growing rapidly. The numbers of data mining consultants, as well as the number of commercial tools available to the "non-expert" user, are also quickly increasing. It is becoming easier than ever to collect datasets and apply data mining tools to them. As more and more non-experts seek to exploit this technology (whether directly via off-the-shelf products or indirectly via consultants) to help with their business, it becomes increasingly important that they understand the underlying assumptions and biases of these tools. One cannot blindly "plug-and-play" in data mining. There are a number of factors to consider before applying data mining to any particular database. This general warning is not new. Many of these issues are well-known by both the data mining experts (Fayyad, et al., 1996) and a growing body of non-expert, data "owners." For instance, the data should be "clean," with consistent values across records and containing as few errors as possible. There should not be a large number of missing or incomplete records or fields. It should be possible to represent the data in the appropriate syntax for the required data mining tool (e.g., attribute/value pairs).

In this paper, we will discuss three specific, but less well-known, issues. Each will be illustrated through real-world experiences. The first is the impact of *data distribution*. Many data mining techniques perform class or group discrimination, and rely on the data containing representative samples of all relevant classes. Sometimes, however, obtaining samples of all classes is surprisingly difficult. The second issue is one of *applicability and data relevance*. High quality data, combined with good data mining tools, does not ensure that the results can be applied to the desired goal. Finally, we will discuss some of the issues associated with using *text* (e.g., narrative fields in reports) in data mining. The current technology cannot fully exploit arbitrary text, but there are certain ways text can be used.

These three issues are not new to the field. Indeed, for many data mining experts, these are important issues which are often well understood. For the non-expert, however, these issues can be subtle or appear deceivingly simple or unimportant. It is tempting to collect a large amount of clean data, massage the representation into the proper format, hand the data tape to the consultant, and expect answers to the most pressing business questions. Although this paper does not describe all of the potential problems one might face, it does describe some important issues, illustrate why they might be problematic, and suggest ways to effectively deal with these situations.

2. TWO EXAMPLES

Our discussion of data distribution, information relevance, and use of text will be illustrated with examples from two current projects. The first involves a joint project with the Center for Advanced Aviation Systems Development (CAASD) in the domain of aviation safety. In this project, one of the primary goals is to help identify and characterize precursors to potentially dangerous situations in the aviation world. One particular way to do this is to mine accident and incident reports involving aircraft for patterns which identify common precursors to dangerous situations. For any type of flight, commercial, cargo, military, or pleasure, accidents (and often less serious incidents) are investigated. A report is filed containing a variety of information such as time of day, type of aircraft, weather, pilot age, and experience. These reports often include the inspector's written summary. One task involves using collections of these reports to try to identify and characterize those situations in which accidents occur. A source of such reports is the National Transportation Safety Board (NTSB).

Another project we are currently working on involves targeting vehicles for law enforcement. In this particular instance, vehicles (mostly passenger vehicles and small trucks) arrive at an inspection stop. At this primary, stop a brief inspection is conducted to decide if further examination is necessary. There is typically a constant flow of cars to be processed, so excessive time cannot be taken. This first inspection typically takes twenty to thirty seconds. If the primary inspector feels it is warranted (and there are any number of reasons which justify this), any vehicle can be pulled out for secondary inspection. This secondary inspection and background check is more thorough. If the driver/vehicle is found to be in violation of the particular laws under consideration, then various information concerning both driver and vehicle is collected and entered into the “violators” database. The goal of this project is to find a way to better profile these violating drivers and vehicles, so that the primary inspectors can more accurately identify likely suspects and send them for secondary inspection.

3. DATA DISTRIBUTION

We first discuss the issue of data distribution. Of particular concern is the situation in which the data lacks certain types of examples. Consider the aviation safety domain. One goal of our project in this domain is to characterize situations which result in accident flights. An obvious source of information is the NTSB’s database of accident reports. Note that this database does not contain records about uneventful flights (the NTSB is an accident investigation agency). That is, the data are unevenly distributed between records of accident flights and records of uneventful flights.

This lack of reports about uneventful flights has important consequences for a significant class of data mining techniques. When given the data containing only accident flights, each of the approaches in this class concludes that all flights contain accidents. Such a hypothesis is clearly incorrect; we know the majority of the flights are uneventful. Also, such a hypothesis is not useful because it does not offer any new insight on how to differentiate the accident flights from the uneventful ones. Furthermore, some of the most popular data mining tools, including decision tree inducers (Quinlan, 1993), neural networks (Rumelhart & McClelland, 1986), and nearest neighbor algorithms (Aha, 1992, Wettschereck, 1994), fall into this class of techniques (i.e., they assume that the absence of uneventful flights in the data implies that they do not exist in the world).

To continue this discussion, it is necessary to first define some terms used in data mining. We say that the *target concept* is that concept we are trying to learn. In the aviation domain, the target concept is accident flights. Consequently, each example of an accident flight (i.e., each accident report in our data base) is called a *member* of the target concept, and each uneventful flight is a *non-member* of the target concept. We have pointed out that the NTSB data do not contain records of uneventful flights. That is, there are no descriptions of non-members of the target concept. The problem of learning to differentiate members from non-members is called a *supervised concept learning problem*. (It is called *supervised* because each example in the data contains a label indicating its membership status for the target concept.) A *supervised concept learner* takes as input a *training sample*. A training sample is a list of examples, labeled as members or non-members, which is assumed to be representative of the whole universe. The supervised concept learner produces *hypotheses* that discriminate the members and non-members in the sample. Many data mining tools use supervised concept learners to find patterns.

Let us say that a supervised concept learner makes the *closed-world assumption* if it assumes the absence of non-members in the data implies that they do not exist in the universe. Why do some of the popular learners make the closed-world assumption? The case of decision tree learners provides a good illustration. These learners partition the training sample into pure sub-samples, containing either all member or all non-members. The partitioning of the training sample drives the rule generation. That is, the learners introduce conditions that define partitions of the training sample; each outcome of a condition represents a different sub-sample. Ultimately, the conditions will become part of the discrimination rules. Unfortunately, if the input sample contains only data which are members of the target class, the training sample is already pure and the decision tree learner has no need to break up the sample further. As a consequence, the rules commit to classifying all new data as members of the target class before conducting any tests. Thus, in the aviation project, all flights would be classified as accident flights, since the learner never saw any uneventful flights. This is not to say that learners employing the closed-world assumption are inappropriate in all, or even most situations. For many problems, when representative data from all the concepts involved is available, these learners are both effective and efficient.

Fortunately, when the data does not contain non-member examples, there are a number of alternative approaches:

Use Different Supervised Concept Learners: There are other supervised concept learners that do not make the closed world assumption. For example, the supervised concept learner AQ has a *uniclass* mode (Stepp, 1979) which takes the members and identifies all the characteristics they share. Because it does not rely on non-member examples, such a learner is better suited for this safety aviation problem than a learner that makes the closed-world assumption.

Obtain Non-Member Examples: In some cases, it may be reasonable to get non-member examples. In the aviation safety problem, it may be possible, although difficult, to collect data about uneventful flights, or obtain data from other sources. Also, it may be possible to use some other existing, but different, data. For example, maintenance logs are kept on all planes. It may be possible to differentiate potential accident flight from uneventful ones using the more popular supervised concept learning programs with this new data.

Use Different Learners: There are learners that can use training examples containing only members of the target concept to produce non-trivial rules. Some *unsupervised* learning methods (Cheeseman, 1988) take the entire set of examples and form *clusters* - groups of data which are similar. (The term unsupervised refers to the fact that the class membership labels in the data are ignored, if they exist at all). In aviation safety, an unsupervised learning tool may take the accident reports and identify that a significant number of the reports involve rainy conditions, young pilots, and small planes (as a hypothetical example). Notice that an unsupervised learner's goal is unlike the goal of AQ with the uniclass mode enabled. Instead of identifying similar characteristics among all the examples, the unsupervised learner separates the set of members into new, similar groups. Of course, there is an obvious drawback to these unsupervised approaches. Knowledge about the high incidence of certain reports does not tell us how to discriminate accident flights and uneventful flights. However, unsupervised learning tools may be able to find information that subsequently helps us find discrimination rules.

Alter the Initial Goal Slightly: It may be possible to choose different, but related, concepts to use in the supervised classification learning. For example, in the aviation data, we might divide the accident flights into two groups: those with pilots under age thirty and those thirty and above. Now, the supervised learning algorithms can be used to differentiate these two groups. One problem with this approach is determining which attribute should be used to assign the new class. That is, why choose “age > 30” rather than “weather is rainy”? A more serious shortcoming is that we are no longer addressing the initial problem of characterizing accident flights versus uneventful flights. However, this type of altered characterization may provide some insight into the accident flights, which could prove useful.

4. APPLICABILITY AND RELEVANCE OF DATA

Even when collected data is of high quality (i.e., clean, few missing values, proper form, etc.) and the data mining algorithms can be successfully run, there still may be a problem of relevance. It must be possible to apply the new information to the situation at hand. For instance, if the data mining produces typical “if...then...” rules, then it must be possible to measure the values of the attributes in the condition (“if” part) of those rules. The information about those conditions must be available at the time the rules will be used. Consider a simple example where the goal is to predict if a dog is likely to bite. Assume data are collected on the internal anatomy of various dogs, and each dog is labeled by its owner as either “likely” or “unlikely” to bite. Further assume that the data mining tools work splendidly, and we discover the following (admittedly contrived) rules:

Rule 1: If the rear molars of the dog are worn, the dog is unlikely to bite.
Rule 2: If the mandibular muscles are over-developed, the dog is likely to bite.

These may seem like excellent rules. However, if faced with a strange, angry dog late at night, these rules would be of little help in deciding whether you are in danger. There are two reasons for this. (1) There is a time constraint in applying the rules. There are only a few seconds to check if these rules apply. (2) Even without such a constraint, the average person probably can’t make judgments about molar wear and muscle development. The lesson here is that just because data are collected about biting (and non-biting) dogs, it does not mean we can predict whether a dog will bite in the situation where it will be most useful.

In the vehicle targeting task described earlier, a similar situation occurred. The initial goal was very specific: develop a set of rules, a profile, that the primary inspectors could use to determine which vehicles to pull out for secondary inspection. As mentioned, much more information is collected concerning actual violators than for those which are just passed through the checkpoint. Thus, the initial goal was to profile likely violation suspects based on the wealth of information about that group. The problem, noticed before any analysis was done, was that the information which would make up the profiles would not be applicable to the desired task. As mentioned, the primary inspectors have only a short amount of time to decide whether a particular vehicle should be pulled out for secondary inspection. During that time, they have access to only superficial information. That is, the primary inspectors don’t have quick access to much of the background knowledge concerning the driver and vehicle. Yet, this is precisely the knowledge collected during seizures and initially chosen to build profiles. Thus, they have no way to apply classification rules which measure features such as “number of other cars owned”, “bad credit history”, or “known to associate with felons” (types of data collected on violation vehicles and drivers).

The problem here is not that the data is “bad”, or even that the data is all from the target concept (see Section 3). The problem is that the data cannot be applied to the initially specified task. How does this situation come about in general? The answer involves a fairly common situation. Often, data mining begins with data which has been previously collected, usually for some other purpose. The assumption is made that since the collected data is in the same general domain as the current problem, it must be usable to solve this problem. As the examples show, this is often not the case. In the vehicle targeting task, the nature of the law enforcement system is such that a great deal of information is collected and recorded on violators. No one ever intended to use this information as a screening tool at stop points. Thus, it is important to understand the purpose for which a set of data was collected. Does it address the current situation directly? Similarly, when data is collected for the specific task at hand, careful thought must go into collecting the relevant data.

There are two primary ways to address this problem of data irrelevancy. The most obvious is to use additional data from another source. It may be that *different* data already exists to address the primary question. For instance, returning to the dogs example, general aggressiveness characteristics for different breeds of dogs have been determined. Using this data, rather than the original data, deciding how likely a dog is to bite is reduced to the problem of determining its breed (often done by quick visual inspection). When the necessary data does not already exist, it may be necessary to collect it. Some of this data collection will likely take place in the vehicle targeting project. In this case, data must be collected which relates directly to the information available to the inspectors at the initial inspection. For example, the demeanor of the driver may be an important feature. Of course, collecting new data may be a very expensive process. First, the proper attributes to collect must be determined. This often involves discussions and interviews with experts in the field. Then, the actual data collection process may be quite costly. It may be that an inordinate amount of manpower is required, or that certain features are difficult to measure.

If additional data cannot be obtained, there is another, often less desirable way to address this issue. It may be possible to alter the initial goals or questions. This will clearly require problem-specific domain expertise to address a few simple questions: Is there another way to address the same issue? Is there another relevant issue that can be addressed directly with this data? In the vehicle targeting domain, we considered using only those attributes which were directly accessible to the inspector. For example, looking at simple statistical patterns for time of day, weather, season, holidays. This is not a very deep analysis and doesn’t quite “profile” likely violators, but it makes progress towards the initial goal. Another alternative is to use the violator database to profile suspects for other situations. It may be that profiles of certain types of violators bear similarities to other criminal types. Perhaps this information can be used elsewhere in law enforcement. Admittedly, this latter solution does not address the initial issue: helping the primary inspectors decide who to pull out for secondary inspection. However, it may not be possible to achieve that goal with this data and the given time constraints. It is important to understand this potential limitation early in the process, before a great deal of time, effort, and money has been invested.

5. COMBINING TEXT AND STRUCTURED DATA

Data mining is most often performed on data that is highly structured. Highly structured data have a finite, well-defined set of possible values, as is most often seen in databases. An example of structured data is a database containing records describing aircraft accidents which includes fields like the make of an airplane and the number of hours flown by the pilot. Another source of valuable yet often unused information is unstructured text. Although more difficult to immediately use than structured data, data mining should make use of these available text resources.

Text is often not used during data mining because it requires a pre-processing step before it can be used by available tools such as decision trees, association rule methods, or clustering. These techniques require structured fields with clearly defined sets of possible values that can be quickly counted and matched. Such techniques sometimes also assume that values are ordered and have well-defined distances between values. Text is not so well behaved. Words may have multiple meanings depending on context (polysemy), multiple words may mean the same thing (synonymy), or may be closely related (hypernymy). These are difficult issues that are not yet totally solved, but useful progress has been made and techniques have been developed so that text can be considered a resource for data mining.

One way to exploit text, borrowed from information retrieval, is to use a vector-space approach. Information retrieval is concerned with methods for efficiently retrieving documents relevant to a given request or *query*. The standard method for doing this is to build weight vectors describing each document and then compare the document vector to the query vector. More specifically, this method first identifies all the unique words in the document collection. Then this list of words is used to build vectors of words and associated weights for the query and each of the documents. Using the simplest weighting method, this vector has a value of 1 at position x when the x^{th} vocabulary word is present in the document; otherwise it has a value of 0. Every document and query is now described by a vector of length equal to the size of the vocabulary. Now each document vector can be compared to every other document by comparing their word vectors. A cosine-similarity measure (which projects one vector along another in each dimension) will then provide a measure of similarity between the two corresponding documents. Surprisingly, although this approach discards the structure in the text and ignores the problems of polysemy and synonymy altogether, it has been found to be a simple, fast baseline for identifying relevant documents.

We used a variant of this vector space approach on the airline safety data to identify similar accidents based on a textual description of the flight history. The narrative description of each accident was represented as a vector and compared to all other narratives using the approach described above. One group of accidents identified by this technique can be described as planes which were ‘veering to the left during takeoff”. The following accident reports were found to be similar in this respect.

-MIA96LA055 - “during takeoff roll he applied normal right rudder to compensate for engine torque. The airplane did not respond to the pilot input and drifted to the left...”
-ANC95LA099- “...veered to the left during the first attempt to take off...”
-ANC95LA041- “...pilot added full power and the airplane veered to the left.”

Identifying this kind of a group would be difficult using fixed fields alone. This technique can also be used to find all previous reports similar to a given accident, or to find records with a certain combination of words. This can be a useful tool for identifying patterns in the flight history of the accident so that the events leading up to different accidents can be more clearly identified.

The information stored in text can be extracted in other ways as well. Feldman (Feldman et. al, 1997) combines a collection of documents and a taxonomy of terms so that maximal word or category associations can be calculated. Although he reported results from newswire data, it could also be used in the airline safety domain to calculate, for example, which class of mechanical malfunctions occurred most often in winter weather.

Another approach very relevant to data mining from text is information extraction (IE). Information extraction is interested in techniques for extracting specific pieces of information from text and is the focus of the ARPA Message Understanding Conference (MUC) (Lehnert and Sundheim, 1991). The biggest problem with IE systems is that they are time-consuming to build and domain specific. To address this problem a number of tools have (and continue to be) developed for learning templates from examples such as CRYSTAL (Soderland, et. al. 1995), RAPIER (Califf, 1997) and AutoSlog(Riloff, 1996). IE tools could be used in the airline safety data to pull out information which is often more complete in the text than in the fixed fields. This work is geared toward filling templates from text alone, but often the text and structured fields overlap in content.

An example of just such an overlap can be found in the NTSB accident and incident records. This data contains structured fields which together allow the investigator to identify human factors as important to the accident. However, it was found that these fields are rarely filled out completely enough to make a classification: 90% of the records which were identified as involving people could only be classified as “unknown”. IE methods could be used to reduce this large unknown rate by pulling information out of the narrative which described if a person in the cockpit made a mistake. Such an approach could make use of a dictionary of synonyms for ‘mistake’ and parser for confirming if the mistake was an action made by the pilot or copilot and not in a sentence describing, for example, the maintenance methods.

Although data mining has primarily concerned itself with structured data, text is a valuable source of information that should not be ignored. Although automatic systems which completely understand the text are a still a long way off, one of the surprising recent results is that simple techniques, which sometimes completely ignore or only partially address the problems of polysemy, synonymy and complex structure of text, still do provide a useful first cut for mining information from text. Useful techniques, such as the vector-space approach and learned templates from information extraction, can allow data miners to make use of the increasing amount of text available on-line.

6. CONCLUSION

We have discussed three data related issues in the context of real-world examples. Section 3 discussed the role of data distribution. If the data contains examples of only a single class, extra work may be involved as some popular types of data mining methods may not be appropriate. Section 4 discussed the applicability and relevance of the data. The data to be mined should have a direct connection to the goal task, and the new information should be directly applicable to the task situation. Finally, Section 5 discussed the role of text in data mining. Although automated understanding of natural language is not available, an increasing number of techniques can be used for exploiting text data.

More broadly, we can summarize these discussions into the following general strategy. At each stage consider the three issues we have discussed: distribution, applicability, and text. Collect appropriate data. Think first about what kind of information is needed and how it will be used. If the data already exist, understand their strengths and limitations as they relate to the task specification and the available data mining techniques. If necessary, consider alternative data sources. It may be possible to augment the existing data with additional data. Finally, if no additional data can be obtained, and the existing data is inadequate for the original task specification, consider altering the objectives.

REFERENCES

Aha, D. (1992). *Tolerating Noisy, Irrelevant and Novel Attributes in Instance-Based Learning Algorithms*. International Journal of Man-Machine Studies 36(1), 267-287.

Califf, Mary E. (1997). *Relational Learning Techniques for Natural Language Information Extraction*. Ph.D Proposal, Department of Computer Sciences, University of Texas at Austin.

Cheeseman, Pl, Kelly, J., Self, M., Stutz, J., Taylor, W., & Freeman, D. (1988). *AUTOCLASS: A Bayesian Classification System*. In Proceedings of the Fifth International Machine Learning Conference, p. 54-64. Ann Arbor, MI: Morgan Kaufmann.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. Menlo Park: AAAI Press.

Feldman, R., Aumann, Y., Amir, A., Zilbertstein, A., Kloesgen, W. (1997). *Maximal Association Rules: A New Tool for Keyword Co-occurrences in Document Collections*. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, (Heckerman, D., Mannila, H., Pregibon, D. and Uthurusamy, R. Eds.), p. 167-170.

Lehnert, W., and Sundheim, B. (1991). *A Performance Evaluation of Text-analysis Technologies*. AI Magazine, 12(3), 81-94..

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Riloff, E. (1996). *Automatically Generating Extraction Patterns from Untagged Text*. Proceedings of the Thirteenth National Conference on AI, p. 1044-1049.

Rumelhart, D.E. & McClelland, J.L. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge: MIT Press.

Soderland, S., Fisher, D., Aseltine, J., and Lehnert, W. (1995). *Crystal: Inducing a Conceptual Dictionary*. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, pp. 1314-1319.

Stepp, R. (1979). *The Uniclass Inductive Program AQ7UN1: Program Implementation and User's Guide*. Report No. 949. Department of Computer Science, University of Illinois, Urbana.

Wettschereck, D. (1994). *A Study of Distance-Based Machine Learning Algorithms*. Ph.D. Thesis, Oregon State University.

THE AUTHORS

Neal Rothleder is a senior staff member of the Artificial Intelligence Technical Center at the MITRE Corporation. His interests include machine learning and data mining, recently focusing on the areas of simplifying classification rules and automatically incorporating domain expertise into the data mining process. Dr. Rothleder received his M.S. and Ph.D. from the University of Michigan. He is a member of the American Association for Artificial Intelligence.
Dr. Rothleder can be reached as follows:
email: neal@mitre.org, phone: (703)983-6909, fax: (703)983-1379

Earl Harris, Jr. is a senior staff member of the Artificial Intelligence Technical Center at the MITRE Corporation. His interests include machine learning and data mining. Mr. Harris received his B.A. degree in Applied Science from Harvard University in 1984, and his M.Sc. from William & Mary in 1990. He has almost completed his Ph.D. at William & Mary.
Mr. Harris can be reached as follows:
email: esharris@mitre.org, phone: (703)983-7170, fax: (703)983-1379

Eric Bloedorn is a senior staff member of the Artificial Intelligence Technical Center at the MITRE Corporation. His interests include machine learning and its application to text classification and data mining. Dr. Bloedorn received his B.A. degree in Physics from Lawrence University in 1989, and his M.Sc. and Ph.D. from George Mason University in 1992 and 1996 respectively. Dr. Bloedorn is a member of the American Association for Artificial Intelligence and the Association for Computing Machinery.
Dr. Bloedorn can be reached as follows:
email: bloedorn@mitre.org, phone: (703)983-5274, fax: (703)983-1379

The address for all the authors is as follows:
The MITRE Corporation