A Dose of Business Intelligence: Data Mining

Joseph Ryan G. Lansangan
University of the Philippines Diliman

With the advancement of data warehousing and processing, large volume and/or different sources of data are now available, such as data on customer profile, transaction details, business processes, and even marketing efforts. Data is processed and summarized into useful information for business strategies. Given the appropriate techniques and tools, companies become proactive on their decisions and/or actions, with insights made for the future using present and/or historical information. Companies value such processes, and hence they continue to gather data, formulate strategies, and make actions, which in turn, become new information and/or yield new business directions. Such cycle of three components—data, question, and decision—remains at the core of today's business intelligence. With the continuous influx of data, questions arise, and hence actions are to be made. Equivalently, with the new directives, information is desired to arrive at certain decisions. And, when actions are made, data comes in and possibly new directions and/or objectives are created.

As an aid to decision-making, crucial to the business intelligence framework (or cycle) is data mining. Data mining is the process of extracting useful (hidden) information or knowledge from large volume of data, commonly implemented on an automated, timely and quick manner as solutions to or support for different analytical queries and/or business problems. Data mining is used to uncover inherent patterns based on historical information, allowing for statistical predictions, characterization and/or classifications of data. Information is then presented in meaningful ways, be it for exploratory reasons (e.g., deep-dive or drill-down analyses) or for modeling purposes. Thus companies with data mining capacity become more forward-looking based on what can be inferred from what information is available. Since data is built from the customers themselves, companies tend to be "customer-centered," and since the processes are carried out to meet certain objectives, companies become "tactically-driven."

Data Mining Techniques

Data mining techniques can be classified into five general areas. First, visual representations techniques are graphical interpretations of complex (and even simple) relationships, which are commonly the "front-end" of other data mining techniques but are also used as "post-hoc" procedures. Data is accessed via specialized views and/or drill-down processes for deeper analyses. Second, variable/feature selection methods are dimension-reduction techniques to summarize data into "relatively fewer" features, commonly used to identify the "more important" information. These are often conducted as data pre-processing, but are also used for index-derivation objectives. Third, segmentation and clustering techniques are used to find groups of "similar" characteristics based on relevant dimensions. Segments or clusters are made based on different similarity (or dissimilarity) measures, the objective of grouping often for profiling purposes, for "targeting" specific segments, or for classifying (of "new" units). Fourth, association rules are used to look for significant relationships and/or sequences among transactions (or events), with the rules based on frequent patterns. Common applications are collaborative filtering, market basket analysis and sequence analysis. Fifth, predictive modeling looks into developing a "model" based on discovered patterns or trends in the data, with the "model" being used to predict future outcome and/or identify impacts of changes in behaviors or activities. Predictive models are commonly used for robust customer valuation (or scoring) and identification (e.g., customers who are most likely to respond to an offer).

The different data mining techniques may address specific objectives, but their essence for a particular company remains the same – to identify and/or understand their customers, gain insights on the company's products and/or services, and take action based on what is presented by or inferred from the data. Visual representations are the most straight-forward, giving deeper perspectives of what the data/information conveys more than what is obvious, using 3-dimensional plots or interactive charts. Feature selection techniques yield interpretable and/or actionable information based on the "best" set/s of variables (relatively fewer than the original set of variables, or combined at fewer dimensions) that capture/s the most from the data. Segments and clusters derived from grouping techniques give deeper comprehension of latent or data-based affinities. Association rules may yield both inexplicable and interpretable rules, but still give knowledge on who the customers are or why customers make transactions (or participate in certain events). Predictive and/or forecasting models are best used to anticipate or forecast patterns or movements, thus the company can decide in a statistical sense (or at calculated risks) using available data.

The Data Mining Process

Different sources in the literature and different data mining software provide different frameworks of the data mining process. But somehow, the data mining process (or any analytical procedure for business intelligence, in this case) can be summarized in three stages – (1) objective and/or data setting (2) data processing and/or analysis, and (3) documentation and execution. These can be further classified as follows – under objective and/or data setting, the company must (a) know the business directives and/or identify specific objectives or queries, (b) then translate the business objectives into analytical objectives, and (c) prepare the data and map out the methodology (if data requirements and/or methods do not suffice to meet the objectives, then the objectives must be re-aligned or the data must be gathered and/or methods must be modified); for data processing, activities include (d) extraction, transformation and loading of data, and (e) analytics proper which includes validation and/or assessment procedures; and finally, activities under documentation and execution include (f) report writing and (g) implementation of decisions/actions.

Note that the discussed stages and/or processes above can be both simplistic and complicated. In the case of setting the business directives, it can be as simple as the Business Intelligence (BI) unit identifying the specific objectives; it can be as not-that-simple as the top management giving general company goals and thus the BI unit works in collaboration with other units (e.g., Marketing unit, Contact center) to come up with specific goals that meet and/or are parallel with the company goals. Translation of the specific objectives into analytical objectives together with data preparation and methodology-sketching are relatively easy tasks, but these become difficult when the company has limited resources (e.g., data sources, software to be used, statistician/s or analyst/s to be engaged, knowledge of methods). The analytics proper has its own simple and complex issues, which basically depend on both the tool/technique and the user. As there are no fixed steps to follow (but standards or best practices remain), analyses are never permanent for a given problem.

How Good Data Mining can Be

To best apply the different data mining techniques, one should not only know what technique is appropriate for a given data, but should always be guided by what the business objective/s is/are. Though it seems that data mining is driven by data, what remains fundamental are (1) the company's motivation or directive – what the company desires to do or needs to address (prior to data mining); and (2) the company's understanding of the results – how the company reacts with the results (during and/or after data mining).

Since data mining entails uncovering hidden information, the discovery process can be complicated, but the effective use of data mining first and foremost lies on the reason/motivation for the conduct of such. Questions and/or objectives must be in place, since these remain as the foundation of all analyses to be made – these questions/objectives essentially define the analyses, taking into account what data or resource is available. As data mining methods can be subjective in nature, approaches may vary but must always be within the scope of the objectives.

Similarly, since the results of data mining are sometimes difficult to understand and appreciate, companies must be able to translate the (mostly quantitative) results into solutions to its business problems or as action-items to meet the business objectives. Levels of interpretability of the results range from easy (e.g., Decision tree models) to difficult (e.g., Neural network models), but the results must always be taken "as is," on the assumption that the data used and the processes made prior to generation of the results are accurate and/or acceptable. Data mining may yield non- or counter-intuitive results, but for as long as the results are extracted from a data using statistically sound processes, the results are (empirically) valid. Rather than challenging such results based on other non-empirical evidences (or, on "similar" studies), such results must be accepted and interpreted in the context the company understands.

Data mining techniques will forever be present – once new data comes in, there will always be something to work on, and hence, innovations are possible. But the importance of an old or new technique must be paralleled with the importance of when/how the technique is used and how the results are interpreted. More often than not, the success of a data mining technique depends not only on the tool or the algorithm/technique, but also on the user's statistical and analytical sense and sensibility.