An Epistemological Approach for Research in Educational Data Mining

Alexander Aristizabal F.
Doctoral student in Education – Universidad Santo Tomas, Bogotá, Colombia.
Assessment Coordinator – Colegio Nueva Granada, Bogotá, Colombia
Postal Address: Calle 135A #9B – 30 Apt. 401 – Bogotá, Colombia
Telephone Number: +57 315 3409150 / Email: alexanderaristizabal@hotmail.com

Corresponding Author:
Alexander Aristizabal F.
Telephone Number: +57 315 3409150 / Email: alexanderaristizabal@hotmail.com

Abstract:
Educational Data Mining is an emerging research field dealing with the finding of hidden patterns in educational data. So far, the research has been focused on the empirical data obtained through various data mining techniques and not necessarily on the actual classroom implications and the knowledge that can be derived from the mined data. In order to expand the body of knowledge about the pedagogical act, an epistemological framework is needed to better understand the nature of knowledge and so avoiding the merely empiro-positivist approach to educational data mining research. After presenting the most common epistemological approach, this paper proposes alternative approaches to construct knowledge as options to develop the research field.
Keywords: Data mining, epistemology, grounded theory, research, knowledge.

1. Introduction

Although the term data mining sounds like a new and modern idea, the concept itself is not. The natural tendency of mankind has always been to make sense of information and to find patterns in data or available information. However, a reality in this society of knowledge is the vast amount of information available and the impossibility of humans to manually process these data and information which makes super computers, with appropriate software, indispensable for performing such work. The main purpose of this article is to present some epistemological approaches that seem to be prevalent in current Educational Data Mining Research and provide additional clues to researchers as to what to consider in order to understand the nature of mined knowledge and so avoid the merely empiro-positivist approach to research in the field. An interpretivist approach

along with the Grounded Theory are suggested frameworks to reach higher grounds of data interpretation and production of usable knowledge by the educational community.

Data mining became a discipline of computer science tracking back its origin to the 80s when it was first used systematically within the community of researchers in this field (Coenen, 2004). Arguably, the approach towards data mining in education and its recognition as a research field begin with the First Symposium of Educational Data Mining in 2005 (Romero, Ventura, Pechenizky & Baker, 2011), the formation of the International Society for Educational Data Mining (EDM) and the publication of the first Journal of Educational Data Mining in 2009 (International Educational Data Mining Society, s. f).

The International Society for Educational Data Mining defines EDM as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings in which they learn. According to Romero, Ventura, Pechenizky & Baker (2011) the primary objective EDM is to use large sets of educational data to improve our understanding of the learning process. It is not surprising that in the couple of decades between the emergence of EDM and present times, the current research in the field has continuously reinforced this emerging area for educational research.

2. Current Research in the Field

Even though, the emphasis of this paper is not to detail a review of the state of art in Educational Data Mining, it is important to mention some reference works in order to know what is being researched and what the methodological and epistemological approaches are available so far.

The first review of the state of the art in EDM was carried-out by Romero & Ventura (Romero & Ventura, 2007) and was published in the journal *Experts Systems with Applications*. This paper initially shows the difference between data mining in general and data mining in the educational arena, particularly in regards to the domain, data types, objectives and techniques, showing that EDM is a growing and independent discipline. For these authors the mined knowledge should facilitate and enhance the learning process as a whole in order to make informed decisions. A little after, Baker & Yacef (2009) published another review showing the history and current trends in the field of Educational Data Mining (EDM), the methodological profile of research in the early years of EDM, the trends and shifts in the research conducted by the community up to that point. Later on, Romero & Ventura (2010) published another review which purpose was to describe the different groups of user, types of educational environments and the data they provide. This particular paper shows the most common educational problems that have been addressed with EDM and delineates future research trends in the field. Mohamad & Tasir (2013) published the latest review which most relevant contribution relies on the presentation of limitations of existing research and some new suggestions for further research.

An overview of EDM research up to this point, indicates that most papers show interest in how to improve and inform the learning and teaching processes by using information mostly from online and computer-mediated learning environments. The mined knowledge has been used by researchers

to propose activities for students, resources, predict academic success, individualize learning, organize contents, and determine correlations between different academic variables and so on.

It is clear that the domains that have most largely contributed to the development of EDM research are computer science through areas such as Machine Learning, Data Mining Techniques, Statistics and Psychometrics. They provide a dimension as to how properly analyze complex study designs, and of course, education and psychology to understand the basics in the teaching and learning processes (Romero, Ventura, Pechenizkiy, & Baker, 2011). Data mining requires programming skills, algorithm development, simulation, statistical models, psychometric analysis and visualization tools, among other skills, which limits its use to very few professions. This means that regular K-12 classroom teachers have had from very little to none contributions to the educational research from the EDM perspective.

In general, and as an opening to the epistemological point of view, research design in EDM has had a strong quantitative approach, which means that empirical data and its interpretation have had a predominant emphasis in the interpretation of variables but very little in the production of theory. So far, there is no indicator of an actual theory of education from educational data mining; what research shows is mostly studies of educational variables from different perspectives and different methods. Hence, an epistemological framework for educational research with data mining is needed to provide researchers with some key elements to understand know knowledge is produced, how science is done and what place in science is being taken the this new emerging field called EDM.

3. Epistemological framework

The following remarks do not to provide researchers with an extensive treatise on epistemology; instead they illustrate some foundations on which current EDM research lays on and how prospective researchers may think of new ways to produce knowledge and expand the frontiers of EDM with new and innovative educational theories. The main premise of this paper is that the epistemological framework is a key component of educational research and provides researchers with the foundations to understand the processes for the advancement of knowledge and the research paths.

When looking at research in EDM, it is very noticeable the value and hierarchy given to data. Researchers find and extract data from learning environments, process them, obtain results, identify patterns and make inferences that are translated into specific actions or recommendations. Therefore, it is not unusual to think that this knowledge is produced in a quantitative research paradigm embedded in a controversial positivist epistemology approach. Positivism as an epistemological point of view was first proposed by Comte (1908) who said that the primary objects are “to generalize our scientific conceptions, and to systematize the art of social life (p. 3)”. For positivists, knowledge is derived from the facts of experience and observation of the natural world. If mined data from educational environments is assumed to be the facts and purely out of this, the researcher derives knowledge, then results are based merely on empirical information. However, these data can be interpreted from different perspectives and the conceptual structure of the researcher plays a vital role in the understanding and further applications of mined data.

It widely expressed that the scientific method is an important way to investigate the natural world, for it keeps research bias-free and relies on objective measures, data or facts. Chalmers (1999) distinguishes three component of the common view of science based on facts:

> "a. Facts are directly given to careful, unprejudiced observers via the senses.
> b. Facts are prior to and independent of theory.
> c. Facts constitute a firm and reliable foundations of scientific knowledge. (p. 4)".

From this perspective, researchers carry out experiments to gather data and reach a full understanding of the world; therefore, concepts and knowledge are created from experience very likely in a hypothetico-deductive fashion. For this approach, "the generation of new knowledge claims is irrelevant for their justification: new knowledge claims are justified by deriving testable consequences from them and by comparing these with observation" (Meheus, 1999). In other words, for a hypothetico-deductive (H-D) confirmation of the evidence should consist of successful predictions that deductively follow from the hypothesis under test (Sprenger, 2011). Educational Data Mining by definition involves the process of discovery of information from data sets and uses processes such as inferences, predictions and correlations, among others, to identify hidden patterns in academic data. This research attempt by discovery seems to be purely positivist and is supported by the fact that research findings indicate that sometimes data is mined with no specific objective in order to find hidden patterns and some other times to test a particular hypothesis. Therefore, if researches reach knowledge from this perspective it seems very likely they are working within a positivist paradigm where data is an objective reflection of the natural phenomena.

Positivist research has had profound criticisms, particularly in the understanding of causal mechanisms. Chi (1998) says: "While [positivism] allows the researcher to discover whether two or more phenomena are linked consistently, it does not explain why the link exists (p. 167)". Hence, the interpretation of quantitative data coming from Data Mining Techniques must go beyond the merely quantitative analysis. In other words, educational research should also include the reasons behind those relationships from the researcher's point of view, making of the interpretation an important component in the analysis of data mining results. It is not casual that through interpretation the…

> Researcher is able to go beyond "what" has occurred to see "how" it has happened. The researcher does this by paying attention to the actors' stated reasons for their behavior, trying to figure out how behavior and belief match, and looking for ways in which those beliefs and behaviors are echoed in other specific practices (Chi, 1998, p. 167).

It is usual that the most suitable way to look at the actors in order to have information for an interpretive approach is through interaction in a face-to-face learning environment. So far, most of the data mining research is done with information from logs and online learning environments, not from direct interactions in a regular classroom. Therefore, in an interpretivist paradigm the "attention to the actors" is crucial. EDM must go to a different level where not everything is machine-mediated, but new mechanisms are devised to use the regular classrooms as endless sources of information yet be mined and interpreted by the researchers, generating new educational theories to advance in the teaching and learning processes.

In the interpretivist paradigm generalizability does not define validity of research as it is for the positivism (Chi, 1998). This is of particular importance in educational research where context is preponderant and findings are not necessarily extrapolated or replicated. Variables and dynamics in the classroom change every lesson, every day and exact reproducibility, as in the case of natural sciences, is not viable. Therefore, generalization from the strict positivist perspective is a risky endeavor in educational research. The prediction of performance, clustering or correlation of variables may work only for a particular group at a given time and space, but given the complex nature of human beings, those findings may not apply to a different context.

In general, the pure positivist paradigm brings about some limitations for this kind of research which means that a new approach is needed in order to better support the interpretivist perspective of the researcher. Some authors called the new approach post-positivism, which according to Phillips & Burbules (2000) is a non-foundationalist approach to human knowledge that rejects the view that knowledge is erected on absolutely secure foundations making fallibilism very much unavoidable. In EDM, data should not be assumed as irrefutable truths, they are just the basic components of a constructed reality. A great deal of educational data changes over time, i.e., the grades a student gets in one semester are subject to many internal and external factors, therefore, the prediction of future performance from that data carries a load of uncertainty. The accuracy of prediction where human interactions take place is risky, it is likely it works to a certain degree for sets of people, but the complexity of individuals lowers the likeliness of specific deductions.

Post-positivists see knowledge as conjectural, which in Popper's terms (1962) indicates that scientific statements are nothing but merely hypothesis or guesses which in many cases turn out to be false. For this reason Popper (1980) suggests what he calls the criterion of falsifiability as a foundational method of the empirical sciences. For Popper "theories may be more, or less, severely testable; that is to say, more, or less, easily falsifiable. The degree of their testability is of significance for the selection of theories (p. 95)". In other words, the acceptance and credibility of a scientific hypothesis or theory is measured in terms of how disprovable it can be, consequently, the acceptance of a deduction from data will rely on how it can be proved otherwise.

When researches propose new hypothesis to enhance or strengthen the body of knowledge in a particular field, what they are doing, in Lakatosian terms is adding auxiliary hypotheses to the hard core of the program (Lakatos, 1978). However, these hypotheses as Popper says: "should always be regarded as an attempt to construct a new system; and this new system should then always be judged on the issue of whether it would, if adopted, constitute a real advance in our knowledge of the world (Popper, 1980, p. 62)". Judgment is done in the community of experts, and it is their competency to determine if the contribution from research constitutes an advancement in knowledge or not. So far, EDM is trying to find hidden patterns in sets of educational data, but the actual contribution to the body of knowledge in pedagogy or didactics is under development and should be validated by experts in education. The empirical data shows evidence of some type of associations, implications, possible outcomes, etc., but the actual translation into theory and practice is still a field that requires more work. The translation into actual classroom theory and practice will only happen the moment classroom teachers begin to explore the findings in the EDM field and start to articulate these findings to their practice. It cannot be considered educational if it only serves

certain learning environments, it is educationally impactful only if it transverses classrooms, teachers, students, parents and institutions.

The goal of EDM should not be limited to finding evident or fuzzy relationships between variables and there is an unarguable fact that in face-to-face classroom there is a wealth of data yet to be mined. Current researchers need to find the way to make software and techniques accessible to the regular teacher so these tools begin to expand and have a meaning and application in the day-to-day teaching experiences. The big potential and power of interpretation lays in the classroom teacher that wants to go beyond the evident.

Since the production of theory and consequent applications are of major interest in education, a suggested methodological approach to merge empirical data with theory could be the Grounded Theory (GT) proposed by Glaser & Strauss in 1967. For them the discovery of theory is possible after a systematic acquisition of data; theory that provides relevant predictions, explanations, interpretations and applications. Their basic position is that "generating grounded theory is a way of arriving at theory suited to its supposed uses in contrast to theory generated by logical deduction from a priori assumptions (Glaser & Strauss, 1967, p. 3)". This statement indicates that when a researcher generates a theory from educational data, whose source can be the result of predictions, associations, clustering, etc., this theory has already a pedagogical purpose and is based on real educational data for a particular context in given space and time. Therefore, this "theory should provide clear enough categories and hypotheses so that crucial ones can be verified in present and future research; (p. 3)". Verification, of course, in terms of having new empirical data that falsifies the theory and strengthens the body of knowledge, which according to Glasser & Strauss (1967), will produce a theory that is not easily refutable by more data or another theory. If EDM is about data and then, the GT is about theory from data, therefore, GT constitutes a consistent way to generate new educational theories from the data derived from the mining analysis of educational data with the advantage for the researcher of not needing preconceived hypothesis. A researcher using the Grounded Theory can be amazed by the emerging findings facilitated by the hidden patterns discovered in data. The process itself can be constructed and reconstructed as the research takes place, which means that for the researcher there is a valuable opportunity to discover the phenomena of greatest importance to participants. In other words, the outcome is not necessarily the researcher's objective, but something emerging from all the people involved.

Jones & Alone (2011) compiled a series of attributes and benefits offered by GT which, if extrapolated, can be the same for educational research:

> …. the method's capacity to interpret complex phenomena (Charmaz, 2003), its accommodation of social issues (Glaser & Strauss, 1967), its appropriateness for socially constructed experiences (Charmaz, 2003; Goulding, 1998), it is imperative for emergence (Glaser, 1978; Glaser & Strauss, 1967), its absence from the constraints of a priori knowledge (Glaser, 1978; Glaser & Strauss, 1967), and the method's ability to fit with different types of researchers (Martin & Turner, 1986).

4. Conclusion

This paper, of course, does not propose a unique and only methodology to construct theory, however, the next step in EDM, involves the researchers in the field to work collaboratively with professionals in psychology, pedagogy, neuroscience or any other field interested in education to make more significant use of the data and novel insights from mining techniques. So far, the research field is focused on the findings and techniques and not particularly in the theory and ground-breaking contributions derived from those findings. Therefore, to grow as a strong and sustained educational research field, some new theories must emerge with reasonable, practical and founded applications in the actual teaching and learning process.

References:

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM-Journal of Educational Data Mining, 1(1), 3-17. Chicago.

Chalmers, A. F. (1999). What is this thing called science? 3rd Ed. Queensland: University of Queensland Press.

Chih, A. (1998). Bridging Positivst and Interpretivist Approaches to Qualitative Methods. Policy Studies Journal. 26 (1), 162-168. Recovered from http://faculty.washington.edu/swhiting/pols502/Lin.pdf

Coenen, F. (2004). Data mining: past, present and future. The Knowledge Engineering Review, 26(01), 25-29.

Comte, A. (1908). A general view ofpositivism. Trans. J. H. Bridges. London: Rout-ledge, 379-383.

Glaser, B. & Strauss, A. (1967). The discovery of grounded theory. Strategies for qualitative research. New Jersey: Aldine Transactions

International Data Mining Society. (s.f.). Recovered from http://www.educationaldatamining.org/

Jones, M., & Alony, I. (2011). Guiding the use of Grounded Theory in Doctoral studies–an example from the Australian film industry. International Journal of Doctoral Studies, (6), 2011.

Lakatos, I. (1978). The Methodology of Scientific Research Programmes. New York: Cambridge University Press.

Meheus, J. (1999). The positivists' approach to scientific discovery. Philosophica (64) 2, 81-108. Recovered from http://logica.ugent.be/philosophica/fulltexts/64-6.pdf

Mohamad, S. K., & Tasir, Z. (2013). Educational data mining: A review. Procedia-Social and Behavioral Sciences, 97, 320-324

Phillips, D. C., & Burbules, N. C. (2000). Postpositivism and educational research. Rowman & Littlefield.

Popper, K. (1962). Conjectures and refutations (Vol. 7). London: Routledge and Kegan Paul.

Popper, K. (1980). La lógica de la investigación científica. Madrid: Editorial Tecnos.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2011). Handbook of educational data mining. CRC Press.

Sprenger, J. (2011). Hypothetico-Deductive Confirmation. Philosophy Compass, 6(7), 497-508. Recovered from http://www.laeuferpaar.de/Papers/CompassH-D.pdf