*Literature Review and Results for Motivation*

Author: Blake Conrad

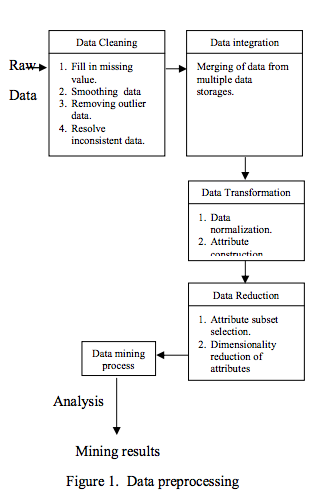
Email: [bmconrad@iupui.edu](mailto:bmconrad@iupui.edu)

Project Advisor: Dr. Murat Dundar

In today’s global economy, population is a factor that effects everyone’s life whether directly or indirectly. An honest correlation to population is the crime associated with it. My capstone project is focused on this precisely; detecting population growth or decay over time, while diagnosing it as healthy, unhealthy, or without noticeable differences. The analysis necessary for this requires data mining for pattern detection. As we will see, there are many options for the way in which one decides to approach the data mining of both categorical and numeric data of all shapes and sizes. For the sake of analyzing *FBI Crime Data from 1979-2014*, we will analyze and discuss the importance of data mining on crime data, the different types of discovery detection and goals scholars have used in the past on similar problems, the differences and uniqueness of this particular project pertaining to the previous applied projects in this area, the different techniques associated with the problem of crime data, researchers best approach at this problem from a historical standpoint, and discussing pros and cons to particular algorithms applied directly to similar data sets of crime.

When analyzing or discussing the importance of data mining in a particular field, it is often coupled with an expert in the field’s domain. For crime data, most data miners work closely with the detectives who will be using the analysis to help assist them in solving crimes or making better crime related decisions. Accomplished faculty members H. Chen and Y. Qin, along with several doctoral candidates at Arizona University collectively have a publishing stating the issues in a concise manner, “Concern about national security has increased significantly since the terrorist attacks on 11 September 2001. The CIA, FBI, and other federal agencies are actively collecting domestic and foreign intelligence to prevent future attacks. These efforts have in turn motivated local authorities to more closely monitor criminal activities in their own jurisdictions” (H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin, M Chau, 2004, 50-51). Without a doubt, data science, data mining, and machine learning are growing in popularity and necessity in almost every industry and agency, especially within governments. Not only is the problem of crime data growing, but the different types of questions being asked about the problem are ever-changing as well. The domain expertise allows for good assumptions in model parameters and what general tendencies should be expected in the data. With respect to my project, I will not have a detective or FBI Agent to work with, so I will have to use some different underlying assumptions to help me obtain good parameters. Noting the different types of questions and expertise required for these types of data, we can consider the uniqueness of different applied projects and their desired data mining goals.

Different questions imply different answers, while results are directly dependent of the answers provided. The faculty at Arizona University state several examples in their publishing, two of which are: unraveling conspiracies based on information on suspects based on geography and detecting cybercrime being found in busy network traffic (H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin, M Chau, 2004, 52-54). Unique problems such as these are difficult in that we are looking for two totally different predictive populations, for the sake of this example, sifting through network traffic to find an internet criminal or raw data tracking geographic locations where issues tend to take place. The publishing states that there are eight different data mining techniques usually found in data mining crime patterns: *Entity Extraction* (identifies particular pattern from data such as text, images, or audio materials), *Clustering Techniques* (used to group items into classes with similar characteristics to maximize or minimize intraclass similarity), *Association Rule Mining* (discovers frequently occurring item sets in a database and presents the patterns as rules), *Sequential Mining* (finds frequently occurring sequences of items over a set of transactions that occurred at different times), *Deviation Detection* (uses specific measures to study data that differs markedly from the rest of the data), *Classification* (finds common properties among different crime entities and organizes them into predefined classes), *String Comparator Techniques* (compares textual fields in pairs of database records and computes the similarity between the records), *Social Network Analysis* (describes the roles of and interactions among nodes in a conceptual network) (H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin, M Chau, 2004, 54-56). Each of the eight different techniques applied to a diverse problem that could be requested pertaining to crime patterns. Without the ability to classify in a supervised manner on my data set, I have narrowed my selection down to *Clustering Techniques*.

 Even after narrowing the different types of problems into the specific category of clustering, there are still several different approaches to the problem and best practices associated with it. Two researchers in Computer Science at Bharathiar University by the names of Revatthy Krishnamurthy and J. Satheesh Kumar have a publishing specifically channeled at my problem labeled, *Survey of Data Mining Techniques on Crime Data Analysis*. These faculties have a UML that shows a good strategy at assessing the type of clustering problem that will involve both categorical and numeric data in *Figure 1* (Krishnamurthy, R., Kumar S. J, 2012, 118). From *Figure 1*, in my project specifically, I plan to follow it almost exactly with some minor deviations and inclusions. I did not intentionally plan to use any type of data reduction, but if time permits this may be a great way to allow for good visualizations to come from high dimensional data. After sifting through the research, this seems to be a good idea. An additional area I am going to apply in the diagram shown is the assessment of centering/not centering the data, including/removing population as a crime statistic, and applying both to the algorithm of choice to determine the best parameters. The diagram in *Figure 1* I am using as a data pre-processing general blue-print, with *Figure 2* being my own personal outlook on the project at large. My project has a good flavor of uniqueness to it, in that I want to rank a city over time as healthy, unhealthy, or stagnant (*Figure 2* demonstrates this as well as the other differentiations of my project to previously studied projects). In addition to simply clustering, I need to see which options are in general good places to start, which options yield the best local optimal solutions, and what the trade-offs might be.

Collectively, it is said that the three best applied algorithms of clustering on detecting crime patterns, listed respectively: *K-means Clustering Algorithm*, *Ak-mode Algorithm*, and *Expectation-Maximization Algorithm* (Krishnamurthy, R., Kumar S. J, 2012, 119). After doing further research in the area of K-means clustering I found a researcher by the name of Shyam Varan Nath from Florida Atlantic University who also works for the Oracle Corporation who has done work directly on this subject. Nath discusses the problem in great detail and decided to use the K-means algorithm, and goes on to state that a big issue was the challenge of finding the best variables to cluster over. Concluding, Nath states their successes and shortcomings, “Our modeling technique was able to identify the crime patterns from a large number of crimes making the job for crime detectives easier,” and “Some of the limitations of our study includes that crime pattern analysis can only help the detective, not replace them” (Nath, S. 2006, 18-22). With the best clustering algorithms generally used for this type of problem is well defined, I will weigh out the pros and cons to my choice from factors such as deadlines, efficiency, and overall practical solutions for this project.

In conclusion, data mining crime patterns has been a well-researched application of data mining. As noted by Andrea Vattani from UC San Diego, “The k-means algorithm is a well-known method for partitioning n points that lie in the d-dimensional space into k clusters. Its main features are simplicity and speed in practice” (Vattani, 324, 2009). Stated by Umale, “The time complexity of the k-means algorithm is where n is the total number of objects, k is the number of clusters, and t is the number of iterations,” making clear expression toward its efficiency (Umale, M, 7). Kumar takes this further in claiming that the complexity of K-means can be brought down even further to if the *Bisecting* version of K-means is applied (Kumar, Karypis, Steinbach, 2). Regarding a comparison of K-means with other algorithms, Savaresi and Boley speak highly on K-means behalf, “K-means is the most popular iterative centroid-based divisive algorithm” (Savaresi, Boley 2011). A limitation to K-means is its clustering model type in that it is limited to its spherical shape (primarily because of the distance definition defined in the algorithm being restricted to Euclidean). The most popular alternative, EM Clustering, has the generally more respected Gaussian models to fit the data. Another advantage to the EM Clustering as stated by Umale, “EM offers many advantages besides having a strong statistical basis and being efficient. One of those advantages is that EM is robust to noisy data and missing information. In fact, EM is intended for incomplete data” (Umale, M, 7). EM Clustering also suffers when its speed is compared to K-means. With EM Clustering being a generalization of K-means, K-means turns out to be faster in processing and defining the clusters. Often, the K-means algorithm is used inside of other higher order algorithms because of its efficiency. After researching the complexity of EM Clustering vs. K-means Clustering, I have chosen to apply K-means to my data set. K-means is a popular algorithm often built into software such as *MATLAB* and *R* which I will be using to run the algorithm on my data and plot results. In addition to the popularity, K-means Clustering is also very fast, so this will help with my frequent and hypothesis testing I want to apply to my project. If time permits I will compare and contrast several algorithms and note the differences. Refer to *Figure 2* for more details regarding the specifics of the project workflow and planning.



Figure 2. Design after data pre-processing

Works Cited

Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., Chau M. (2004). *Crime Data Mining: A General Framework and Some Examples*, 50-56, IEEE. doi: [10.1109/MC.2004.1297301](https://doi.org/10.1109/MC.2004.1297301)

Krishnamurthy, R., Kumar S., J. (2012). *Survey of Data Mining Techniques on Crime Data Analysis*, Volume: 01, Issue: 02, 117-120, Dec 2012, International Journal of Data Mining Techniques and Application, ISSN: 2278-4219

Nath, S. (2006). *Crime Pattern Detection Using Data Mining*. IEEE. doi: [10.1109/WI-IATW.2006.55](https://doi.org/10.1109/WI-IATW.2006.55)

Vattani Andrea. (2009). *K-means Requires Exponentially Many Iterations Even in the Plane*. Proceedings of the twenty-fifth annual symposium on Computational geometry, ACM New York, NY, USA. doi: [10.1145/1542362.1542419](https://doi-org.proxy.ulib.uits.iu.edu/10.1145/1542362.1542419)

Savaresi, M., S., Boley, L., Daniel. (2001). *On the performance of bisecting K-means and PDDP*, Proceedings of the 2001 SIAM International Conference on Data Mining. Doi: [http://dx.doi.org/10.1137/1.9781611972719.5](http://dx.doi.org.proxy.ulib.uits.iu.edu/10.1137/1.9781611972719.5)

Steinbach, M., Karypis, G., Kumar, V. *A Comparison of Document Clustering Techniques*. Retrieved from <https://pdfs.semanticscholar.org/c110/0f525044b2b926f7bd7f407ce7b0157bcfd8.pdf>

Umale, B., M, N. *Overview of K-means and Expection Maximization Algorithm for Document Clustering*. Retrieved from http://research.ijcaonline.org/icquest/number1/icquest1510.pdf