*Literature Review and Results for Motivation*

Author: Blake Conrad

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

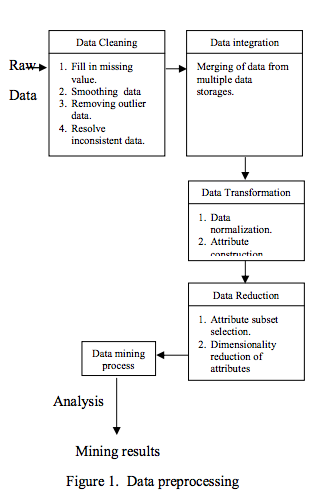
Project Advisor: Dr. Murat Dundar

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 previously applied data mining goals, the different techniques associated with the problem of crime data, researchers best approach at this problem from a historical standpoint, and discussing successes and shortcomings 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 area associated with the data mining being done. 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 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, with the FBI as no exclusion. Not only is the problem of crime data growing, but the different types of questions being asked about the problem are diverse. Noting the different types of questions, we can consider the uniqueness of different problems pertaining to previously applied 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, the network traffic finding 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 data 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 of how to attack such a problem. 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 plot 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. One area I am not going to apply from the diagram shown, is in the data reduction, for the time line associated with the project I don’t believe it would be practical to include this piece and make sense of its impact on the analysis. An additional area I am going to apply in the diagram shown, is the assessment of centering and not centering the data, including/removing population as a crime statistic, and applying both to the algorithm of choice to determine the best parameters.

Collectively, they state 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. In conclusion 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).

In conclusion, data mining crime patterns has been a well-researched application in the field 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). Regarding a comparison of K-means with another algorithm, 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 it’s clustering model type in that it is limited to its spherical shape (primarily because of the distance definition defined in the algorithm). The most studied alternative, EM Clustering, which ails with respect to speed, has a much better Gaussian models to fit the data. K-means turns out to be a particular instance of the generalized EM Clustering, but has a much faster speed for processing the data. Often, the K-means algorithm is used inside of other higher order algorithms because of its effectiveness.

My project advisor agreed on behalf of K-means being sufficient enough for the sake of the clustering in this project where Euclidean distance is a good metric. My project in particular has slightly more detail than the any of the research I have found on this exact type of issue; clustering crime patterns. I have incorporated the idea of a ranking system to introduce another layer of complexity on top of the K-means clustering after processing, diagnosing healthy or unhealthy growth over time. This yields both interesting problems and conclusions, as I can only lean on best practices and research to a certain extent. There exists plenty of overlap between my design layout in *Figure 2* and the design layout of Krishnamurthy and Kumar in *Figure 1*, so I have a well-tested design pattern to leverage for this project. Determining is a problem that *can* be found systematically, but can only be interpreted by domain expertise. My project advisor stated that a level of creativity is often needed for an interpretation without domain expertise, so the real issue is not the effectiveness of K-means, but rather the determining of a good parameter . With that, I will be leveraging popular public data to determine criteria that will yield a sufficient . With K-means we have the benefit of being fast and effective for the distance metric as Euclidean with respect to a mean . With speed becoming an issue, the research collected here has led to the application and interpretation of the project’s main algorithm being K-means.



Figure 2. Design after data pre-processing

Works Cited

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

Krishnamurthy, R., Kumar S. J., *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. *Crime Pattern Detection Using Data Mining*, 18-22 Dec. 2006. IEEE. doi: [10.1109/WI-IATW.2006.55](https://doi.org/10.1109/WI-IATW.2006.55).

Vattani Andrea. *K-means Requires Exponentially Many Iterations Even in the Plane*. 324-332, Jun 2009, Aarhus Denmark. 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. *On the performance of bisecting K-means and PDDP*, 2001 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)