Classification of CKD Cases Using MultiVariate K-Means Clustering

Abhinandan Dubey

July 25, 2015

Abstract

The automated detection of diseases using Machine Learning Techniques has become a key research area lately. Although the computational complexity involved in analyzing a huge data set can be extremely high, nonetheless the merits of getting a desired result surely counts for the complexity involved in the task. In this paper we adopt the K-Means Clustering Algorithm with a single mean vector of centroids, to classify and make clusters of varying probability of likeliness of suspect being prone to CKD. The results are obtained from a Real Case Data-Set from UCI Machine Learning Repository.

1 Introduction

The notion of classifying things has always been a critical question in course of human belief and thought. Classification is deliberated as an instance of supervised learning – the learning process in which a training set of properly identified observations is given as input. The equivalent unsupervised procedure is identified as clustering, and comprises of grouping raw data into "clusters" based on some measure of inherent similarity or distance.

The enormous advance of the amount of biological data available has raised a grim question of being classified, managed effectively and to be transformed from raw data to meaningful information. The emergence of this colossal amount of calls into question the paradigms of modern computation. It seeks an answer towards getting meaningful results out of it, keeping a distinct reasoning for underlying algorithms. Machine Learning surely stands to capture a major fraction of the problem and thus accounts for the latest progress in the field of bioinformatics, computational biology and application of machine learning methods on prominent problems in human biology and behaviour. [1]

The algorithms and mathematical techniques allow us to go beyond a mere depiction of the data and make offers logical results in the form of mathematically testable models. The notions of supervised and unsupervised learning makes this process easy and comprehensible. By simplifying abstraction that institutes a model, we are be able to obtain statistical predictions of a system.

1.1 Unsupervised Learning

The core objective of Unsupervised Learning is for the program or the system to find "patterns" or what we specifically call "clusters" within a given set of data. The data can be matched to a known set of results which can be even used as a classification technique, after analyzing the results obtained from the clustering algorithm.

A major problem of Unsupervised Learning is to give an accurate domain of the clusters and find their centroids. For this, we use various clustering algorithms

1.2 Clustering Analysis

Cluster analysis or simply put forward, "clustering" is the process of grouping a set of elements or data in such a way that elements in the same group (referred to as a cluster) are in share something common to each other than to those in other groups ("clusters"). The task of clustering is often accomplished by clustering algorithms which seek to optimize the data clustering. The process of clustering is not only a main task of exploratory data mining, but it is a very common technique used in statistical data analysis, and many other distinct fields, such as pattern recognition, genetic algorithms, image analysis, information retrieval, and bioinformatics.

Cluster models can be constructed on the basis of some predefined underlying criteria. Several models have been proposed. Some of them are as follows;

- Connectivity models: The models involving euclidean distance connectivity. for instance, hierarchical clustering constructs models based on distance connectivity.
- Centroid models: The centroid models are the most commonly used ones. Their convenience and simplicity makes it feasible for the programmer to deal with a large data set. This is the model we have adopted in this paper.
- Distribution models: In these models, Clusters are modeled on the basis of statistical distributions, for example multivariate normal distributions constructed by the Expectation-maximization algorithm.
- Density models: They explore the connected dense data regions in the raw data space.
- Subspace models: These models are widely used in Two-clustering or Bi-clustering models. Cluster members and attributes that are relevant are used for constructing the clusters.
- Group models: When grouping information is the only prominent output of a clustering system, it is called as a Group Model
- Graph-based models: It involves "a clique" which is a subset of various nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Quasi-cliques also exist, such as in HCS-Clustering Algorithm.

1.3 K-Means Clustering or Lloyd's Algorithm

K-Means Algorithm, also called Lloyd's Algorithm is one of the most simplest clustering algorithms that provide effective results in Unsupervised Learning. The "K" refers to the number of clusters, or centroids in which data set has in which data set has to be classified. As discussed above, the model is based upon centroid clustering. These centroids are calculated after a series of calculations which further optimze their location. A relatively large distance between these centroid coordinates is more favourable. The next step is to map each point to a distinct cluster to which its distance is minimum. Given a set of observations (x1, x2, ..., xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into $k(\leq n)$ sets $S = S1, S2, \ldots, Sk$ so as to minimize the within-cluster sum of squares (WCSS). Hence, its objective is to optimize:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

where μ_i is the mean of points in S_i.

Algorithm 1. *(K-Means Clustering Algorithm)*

Given a first set of k means $m_1(1), ..., m_k(1)$, the algorithm continues by alternating between two steps: [2]

Assignment step: Allocate each observation to the cluster whose mean produces the minimum within-cluster sum of squares. Often, the within-cluster sum of squares is referred to as WCSS, a prominent aim of clustering algorithms. The legitimacy of the statement that this mean is actually minimal can be satisfied by the notion that it is calculated by Euclidean distance formula.

$$S_i^{(t)} = \big\{x_p : \big\|x_p - m_i^{(t)}\big\|^2 \leq \big\|x_p - m_j^{(t)}\big\|^2 \ \forall j, 1 \leq j \leq k\big\},$$

where each x_p is allocated to exactly one $S^{(t)}$, even if it could be allocated to two or more of them.

Update step: This step is an important one as it determines the centroid values of all the clusters. The new means are calculated to be assigned to the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Subsequently, the arithmetic mean is a least-squares estimator thus it also reduces the within-cluster sum of squares (WCSS) objective. Because both steps optimize and achieve the minimal WCSS, and as there only exists a finite number of such cluster partitionings, the algorithm must converge to a (local) optimum. However, the algorithm provides no guarantee that a global optimum is found.

1.4 Chronic Kidney Disease

Chronic kidney disease (CKD), involves a continuous loss in renal function which may remain progressive over several months, or if untreated, even years. Also

Attribute	Quantity	Value-Type
Blood Pressure	mm/Hg	Numerical
Serum Creatinine	mgs/dl	Numerical
Packed Cell Volume	Percent	Numerical
Hypertension Factor	Number	Numerical
Anemia Factor	Number	Numerical

Table 1: Attributes In The Training Set To Calculate L-factor.

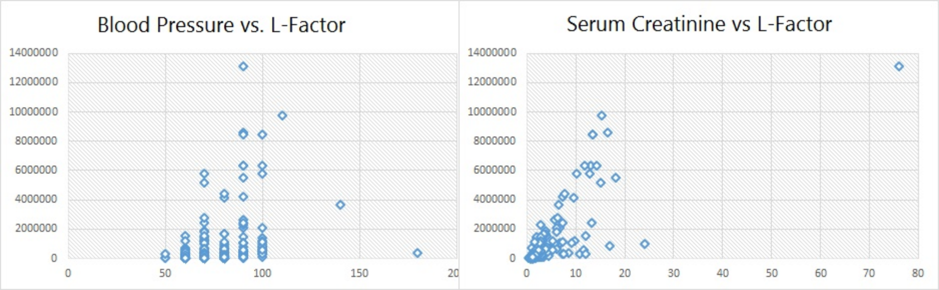
known as chronic renal disease, the symptoms which contribute collectively towards worsening of kidney function are not explicit, and might include feeling unwell for longer periods of time and experiencing a reduced appetite. Often, CKD is diagnosed as a result of screening of people identified to be at risk of kidney problems, for example those with high blood pressure or diabetes and those with a blood relative with CKD. The disease has posed several problems. The major factors involving are Blood Pressure, Sugar Levels, and Anaemia with unusual Creatinine levels. (We thus take into these prominent factors to calculate "L-factor" which will be defined further. CKD may also be identified when it leads to one of its known complications, such as anaemia, cardiovascular disease, or pericarditis. Hypertension is also a known complication of CKD. It is distinguished from acute kidney disease in that the decrease in kidney function must be existent for over 3 months

We have chosen only some of the attributes such as blood pressure, sugar, hypertension, reatinine levels and Anemia to calculate L-factor which is the input to our clustering algorithm. According to a survey by Joseph Coresh & Astor, The prevalence of CKD in the US adult population was 11% (19.2 million). An estimated 5.9 million individuals (3.3%) had stage-1 (persistent albuminuria with a normal GFR), 5.3 million (3.0%) had stage 2 (persistent albuminuria with a GFR of 60 to 89 mL/min/1.73 m2), 7.6 million (4.3%) had stage 3 (GFR, 30 to 59 mL/min/1.73 m2), 400,000 individuals (0.2%) had stage 4 (GFR, 15 to 29 mL/min/1.73 m2), and 300,000 individuals (0.2%) had stage 5, or kidney failure. Aside from hypertension and diabetes, age is a key predictor of CKD, and % of individuals older than 65 years without hypertension or diabetes had stage 3 or worse CKD. [3]

1.5 Preprocessing The Training Set - L-factor Calculation

The dataset we use can be used to predict the chronic kidney disease and it is collected from various Indian hospitals in nearly 2 months of period. The original dataset contains more entries, however, since some entries were missing substantial amount of information, we have excluded them from our consideration in the training set. Moreover, it is obvious that L-factor is not a very solid measure of likeliness. We have taken only a few attributes in L-factor calculation.

To get a more clear and concise idea of a variety of attributes in the training set, we do a bit of pre-processing before we apply the algorithm presented in Section (1.3) above.



(a) Variation of Blood Pressure with L-factor (b) Variation of Serum Creatinine with L-factor

Figure 1: Variation of L-factor

Definition 1. *We define L-factor of a case as follows:*

$$L - factor = b \times s_c \times \pi \times f_1 \times f_2$$

$$\begin{aligned} \textit{where } b &= \textit{Blood Pressure in mm/Hg,} \\ s_c &= \textit{Serum Creatinine in mgs/dl} \\ \pi &= \textit{Packed Cell Volume} \\ f_1 &= \textit{Hypertension Factor (Present = 15, Absent = 4)} \\ f_2 &= \textit{Anemia Factor (Present = 15, Absent = 4)} \end{aligned}$$

f_1 is defined suitably to account for the effect of Hypertension symptoms in CKD. Similarly, f_2 is defined to account for the effect of Anemia in CKD.

The following scatter plots in Figure (1) show the variation of L-factor of 308 suspects from the our Training Set.

2 Processing & Analysis

Let us consider the inputs of our algorithm presented in Section (1.3) Having calculated the L-factors, we have a relatively simple input. $X = 308 \times 1$ Vector with values of L-factor calculated in Section (1.5), and dimension d coincides with unity in our case.

Now, we apply K-Means algorithm to the obtained L-factor values of all 308 suspects. We take K=3 to obtain three different clusters.

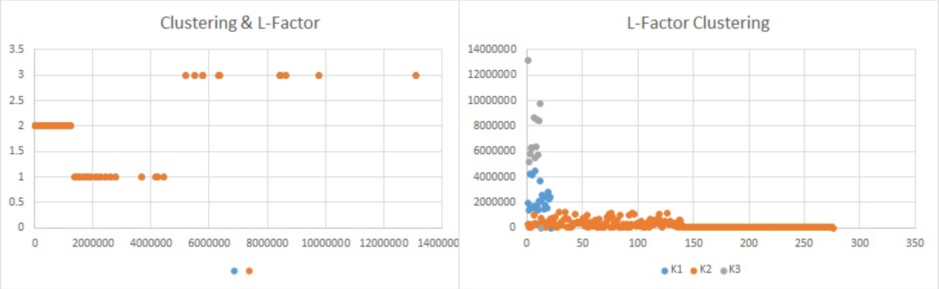
We observed that a unique cluster of values from the training set constituted in suspect being 100% prone to CKD. Though the validity of such calculation is vulnerable to scientific debate, but the results completely shore up that a CKD case is likely to fall in one of the highly prone clusters.

We obtain three clusters with centroids given in Table (2)

The three clusters thus obtained can be plotted against their corresponding L-factor value. This has been shown in Figure 2.

Cluster	Centroid	Number of Values
K1	2336390.714285714	21
K2	214037.5927272727	275
K3	7476675.0	12

Table 2: Clusters Obtained from K-Means Clustering



(a) Clustering & L-factor

(b) L-factor Clustering

Figure 2: Clustering and L-factor

3 Results & Correctness of the Method

Upon comparing the results of K-Means Clustering with the actual (known) result in the training set, we observe that the suspects falling in clusters K1 or K3 are surely suffering from CKD. The result cannot though prove firmly the cases of the K2 cluster, which seem to be distributed in the two classes (CKD/Non-CKD). The probability of a suspect lying in K2 cluster to fall in the class of CKD is 0.50545, which implies that the suspect cannot be classified by our L-factor classifier. However, suspects from clusters K1 & K3 were found to be falling in CKD class with full probability.

References

[1] Ethem Alpaydin. Introduction to machine learning. *MIT Press*, pages 9–12, 2010.

[2] David Mackay. Chapter 20. an example inference task: Clustering. *Information Theory, Inference and Learning Algorithms. Cambridge University Press.*, pages 284–292, 2003.

[3] Josef Coresh, Brad C. Astor, Tom Greene, Garabed Eknoyan, and Andrew S. Levey. Prevalence of chronic kidney disease and decreased kidney function in the adult US population: Third national health and nutrition examination survey. 41(1):1–12.

[4] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[5] Andrew S. Levey, Josef Coresh, Ethan Balk, Annamaria T. Kausz, Adeera Levin, Michael W. Steffes, Ronald J. Hogg, Ronald D. Perrone, Joseph Lau,

and Garabed Eknoyan. National kidney foundation practice guidelines for chronic kidney disease: Evaluation, classification, and stratification. *Annals of Internal Medicine*, 139(2):137–147, 2003.

[6] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. 28:100–108.

[7] L.Jerlin Rubini(Research Scholar) Dr.P.Soundarapandian.M.D. D.M (Senior Consultant Nephrologist) and (Alagappa University) Dr.P.Eswaran. Chronic kidney disease data set - uci machine learning repository. 2015.

[8] L.Jerlin Rubini Dr.P.Soundarapandian. Chronic kidney disease data set - uci machine learning repository, 2015.

[9] M. Lichman. UCI machine learning repository, 2013.