IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

K-Means Clustering for Horse Colic Data

Sandeep Godara

Abstract: Equine colic is a relatively common disorder of the digestive system. Although the term colic, in the true definition of the word, simply means "abdominal pain," the term in horses refers to a condition of severe abdominal discomfort characterized by pawing, rolling, and sometimes the inability to defecate. Clustering is one of the unsupervised learning method in which a set of essentials is separated into uniform groups. The k-means method is one of the most widely used clustering techniques for various applications Cluster analysis for Horse colic data sets has proved to be a useful tool for identifying biologically relevant groupings of genes and samples. In this paper the K-means algorithm is used for clustering Horse Colic Data Set.

Keywords: Data mining, Clustering, K-Means Clustering, Horse colic.

I. INTRODUCTION

Colic in horses is defined as abdominal pain, but it is a clinical sign rather than a diagnosis. The term colic can encompass all forms of gastrointestinal conditions which cause pain as well as other causes of abdominal pain not involving the gastrointestinal tract. The most common forms of colic are gastrointestinal in nature and are most often related to colonic disturbance [1]. There are a variety of different causes of colic, some of which can prove fatal without surgical intervention. Colic surgery is usually an expensive procedure as it is major abdominal surgery, often with intensive aftercare. Different images of horse colic suffering is shown in Figure 1.

Fig 1: Result of the cluster analysis

Among domesticated horses, colic is the leading cause of premature death. The incidence of colic in the general horse population has been estimated between 4[1] and 10[2] percent over the course of their lifetime. Clinical signs of colic generally require treatment by a veterinarian. Colic can be divided broadly into several categories [2]:

- excessive gas accumulation in the intestine (gas colic)
- simple obstruction
- strangulating obstruction
- non-strangulating infarction
- inflammation of the gastrointestinal tract (enteritis, colitis) or the peritoneum (peritonitis)
- ulceration of the gastrointestinal mucosa

Among the general equine population four to 10 out of every 100 horses experience a colic episode during their lifetime; of these, 1-2% require surgical treatment (2006AAEP Proceedings). In 35 years of performing colic surgeries, David Freeman, MVB, PhD, Dipl. ACVS, professor of large animal surgery at the University of Florida's College of Veterinary Medicine, has noticed some trends: For example. 33-50% of referred colic cases now go on to surgery while most of the rest are resolved with medical treatment. On the other hand, Eric Mueller, DVM, PhD, Dipl. ACVS, professor and director of equine programs at the University of Georgia's College of Veterinary Medicine, reports that 25-30% of horses admitted to the university clinic for colic require surgery while the rest are treated with medical therapy. "In 5-10% of cases a horse is humanely euthanized because of a poor prognosis or economical considerations,"[1][3].

Clustering is a method of un-supervisory learning and a common technique for data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering is a division of data into groups of similar objects. Clustering can be used to group the patern which are suffering from disease or not suffering. Each group, called a cluster, consists of objects that are similar between them-selves and dissimilar compared to objects of other groups. The following diagram in Figure 2 represents the clustering process [14]:

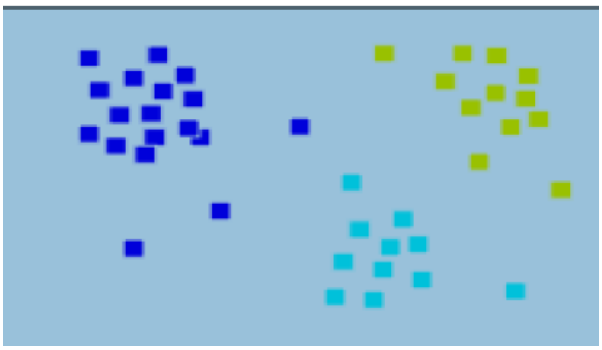


Fig 2: Result of the cluster analysis

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

There are several commonly used clustering algorithms, such as K-means, Density based, Hierarchical and so on. Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. As its well known that each clustering algorithm, sometimes even the same clustering algorithm applied several times on initial dataset, can result in different partitions. The selection of subset of attributes for Clustering as well as number of clusters also effects performance of clustering. Osama Abu Abba et al. Paper is intended to study and compare different data clustering algorithms. The algorithms under investigation are: k-means algorithm, hierarchical clustering algorithm, self-organizing maps algorithm and expectation maximization clustering algorithm. Jain et al. Presented an overview of pattern clustering methods from a statistical pattern recognition perspective. D.Napoleon et al. present that K-means clustering algorithm often does not work well for high dimension. Adil M. Bagirov et al. Developed a new version of the global k-means algorithm, the modified global k-means algorithm [15]. Clustering algorithms based on global optimization techniques are not applicable to even relatively large data set. Narendra Sharma et al. presents the study of various clustering algorithms. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. WEKA tool was used for comparisons[8][9][10].

Bradley et al. present a technique for initializing the K-means algorithm. They begain by randomly breaking the data into10, or so, subsets. They then performed a K-means clustering on each of the10 subsets, all starting at the same set of initial seeds, which are chosen randomly. The result of the 10 runs is 10K centre points. These 10K points were then themselves input to the K-means algorithm and the algorithm run 10 times, each of the 10 runs initialized using the K final centroid locations from one of the 10 subset runs. The resulting K centre locations from this run are used to initialize the K-means algorithm for the entire dataset.

Khan et al. present an algorithm to compute initial cluster centers for K-means clustering. This algorithm is based on two observations that some of the patterns are very similar to each other and that is why they have same cluster membership irrespective to the choice of initial cluster centers. Also, an individual attribute may provide some information about initial cluster center[13][14].

Clustering of a set forms a partition of its elements chosen to minimize some measure of dissimilarity between members of the same cluster .Clustering algorithms are often useful in various fields like data mining, pattern recognition, learning theory etc[11][12].

II. K-MEANS CLUSTERING ALGORITHM

The K-means algorithm:

K-means algorithm follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main idea is to define k centroids, one for each cluster. The simple K-means algorithm chooses the centroid randomly from the data set. The next step is to take each data belonging to a given data set and associate it to the nearest centroid. The K-means clustering partitions a data set by minimizing a sum of squares cost function[5].

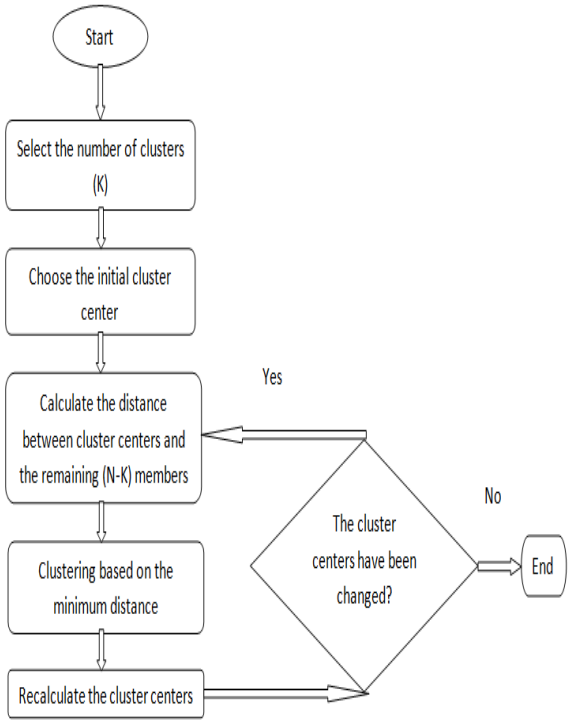


Figure 3: K-means algorithm steps [9].

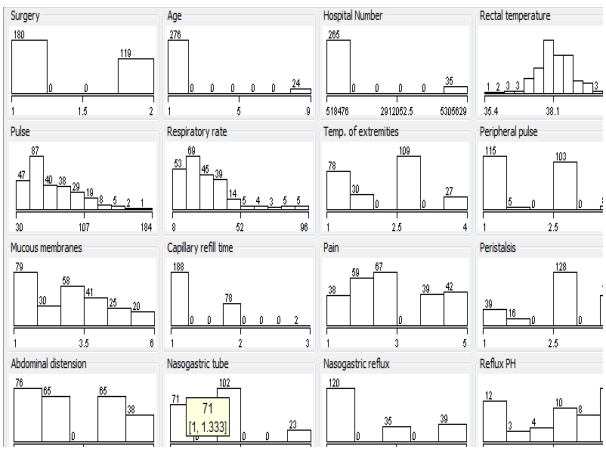
III. EXPERIMENTAL SETUP & PERFORMANCE METRICS

A.Data sets Used:

Horse Colic data set:

This data set is downloaded from archive.ics.uci.edu/ml/ datasets.html and has 37 attributes and 300 instances.

Figure 4 shows Probability distribution function off all 37 attributes. WEKA 3.7 is used for this clustering research work.



ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

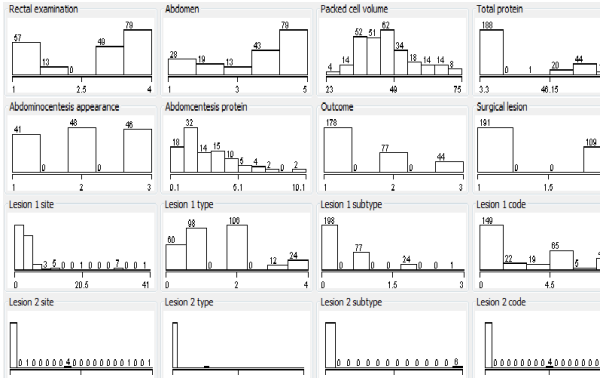


Figure 4: Pdf for various attributes

B. Comparative Analysis:

The K-Means, is applied on Horse Colicl data set. and their results are obtained in terms of TP ,FP, Precision, Recall, Fmeasure and ROC . The upper left cell shows the number of samples classifies as true while they were true (i.e., TP), and the lower right cell shows the number of samples classified as false while they were actually false (i.e., TN). The other two cells show the number of samples misclassified. Particularly, the upper right cell showing the number of samples classified as false while they actually were true (i.e., FN), and the lower left cell showing the number of samples classified as true while they actually were false (i.e., FP).

```
=== Confusion Matrix ===

   a   b   <-- classified as
  29  70 |   a = 1
  72 129 |   b = 2
```

Figure 5: shows Confusion Matrix for Horse Colic Dataset.

```
Correctly Classified Instances         158               52.6667 %
Incorrectly Classified Instances       142               47.3333 %
Kappa statistic                          -0.0649
Mean absolute error                       0.4733
Root mean squared error                   0.688
Relative absolute error                 106.9324 %
Root relative squared error             146.3069 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.293    0.358      0.287     0.293     0.29       0.467    1
                 0.642    0.707      0.648     0.642     0.645      0.467    2
Weighted Avg.    0.527    0.592      0.529     0.527     0.528      0.467
```

Figure 6: shows TP, FP, Precision, Recall, Fmeasure and ROC for Horse Colic data set.

The given Figure 6 shows respective results. The Figure5 shows the confusion matrix of Horse Colic data.

Finally, the generated results by K-Means outperform K-Means and in terms of, accuracy TP, FP, Precision, Recall, Fmeasure and ROC are shown in Figure6 and has attained 52.667% accuracy. So above models have good predictive capabilities for for Horse Colic Dataset.

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) = 52.667$$

TABLE1: CONFUSION MATRIX

	Classified as Healthy	Classified as not healthy
Actual Healthy	TP	FN
Actual not healthy	FP	TN

TP and FP both are high which indicates good clustering capability of K-Means. FP and FN are low which shows that misclassification rate is low. Figure 7 shows clusters made by K-means for Horse Colic Dataset. There are two clusters shown as blue and red. There is good sepration between these two which indicates good capability of K-Means.

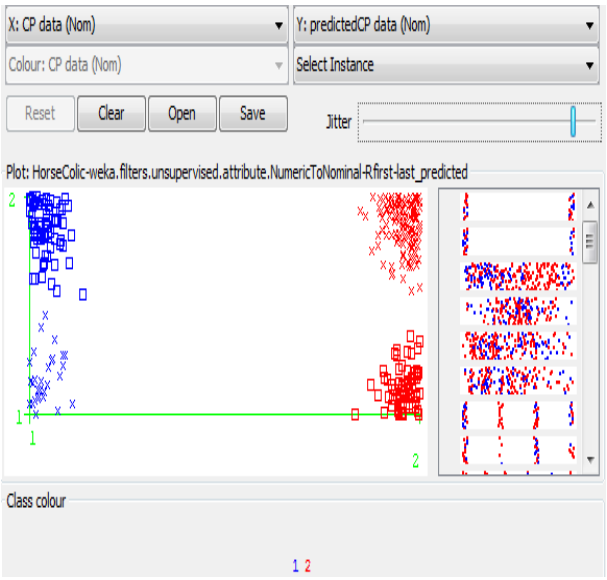


Figure 7: shows clusters for Horse Colic Dataset.

IV. CONCLUSION

We have clustered Horse colic dataset using K-means algorithm and the experimental results have showed 52.66% accuracy. Tp and FP values are high which indicates high rate of classification. The proposed work can also be explored by use of various filtering algorithm for data preprocessing. Further we will apply certain constraints on K-Means algorithm which will not only improve its cluster quality but also its efficiency and clustering capabilities.

DOI 10.17148/IJARCCE.2016.5927

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

REFERENCES

[1] Larson, Erica. "Equine Postoperative Ileus Insights". www.thehorse.com. The Horse. Retrieved 4 July 2014.
[2] Loving, Nancy. "Equine Colic Management and Long-Term Survival". www.thehorse.com. The Horse. Retrieved 4 July 2014.
[3] Moore, James. "Gas Colic". www.thehorse.com. The Horse. Retrieved 4 July 2014. A.K. Jain, M.N. Murty, P.J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, vol. 31, no. 3, pp. 264-323,Sep. 1999.
[4] D.Napoleon,S.Pavalakodi,"A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set," International Journal of Computer Applications (0975– 8887),vol. 13, no.7, pp.41-46, Jan 2011.
[5] Adil M. Bagirov Karim Mardaneh," Modified global k-means algorithm for clustering in gene expression data sets", Workshop on Intelligent Systems for Bioinformatics (WISB2006), Hobart, Australia,vol 73,2006.
[6] Christopher M. Bishop and Michael E. Tipping, "A Hierarchical Latent Variable Model For Data Visualization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp.281-293, Mar. 1998.
[7] M.Awad, H. Pomares, I.Rojas, Member, IEEE," Enhanced Clustering Technique in RBF Neural Networkfor Function Approximation",Mathematical and Computer Modelling,Vol 55,Issues3-4, pp. 286-302, Feb. 2012.
[8] Fasahat Ullah Siddiqui and Nor Ashidi Mat Isa,"Enhanced Moving K-Means (EMKM) Algorithmfor Image Segmentation,"IEEE, pp.833-841.
[9] Mu-Chun Su and Chien-Hsing Chou, "A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp.674-680, Jun. 2001.
[10] TapasKanungo,David M.Mount ,NathanS. Netanyahu, Christine D.Piatko, Ruth Silverman, and Angela Y.Wu,"An Efficientk-Mean Clustering Algorithm: Analysisan Implementation ,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-891,Jul 2002.
[11] Sunila, Prabhat, and Nirmal Godara. "Decision support system for cardiovascular heart disease diagnosis using improved multilayer perceptron." *International Journal of Computer Applications* 45.8 (2012).
[12] Sunila Godara, Amita Verma, "Analysis of Various Clustering Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-3, Issue-1, Page 186-189 and June 2013,.
[13] Sunila Godara,., and R. Yadav. "Performance analysis of clustering algorithms for character recognition using Weka tool." *International Journal of Advanced Computer and Mathematical Sciences* 4.1 (2013): 119-123.
[14] Sunila G, Rishipal S. "Machine Learning For Medical Decision Support Systems (MDSS): A review.", International Journal of Applied Engineering Research. 2015; 10(13):32864–73.
[15] Sunila Godara, Rishipal Singh, "Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis", Indian Journal of Science and Technology, vol. 910, 2016.