Non-Technical Losses Identification Using K-Means Algorithm

Kalyan Pal

Dept. of Electrical Engineering, ABES Institute of Technology, Ghaziabad, U.P., India

Bhavesh Kumar Chauhan

Dept. of Electrical Engineering, ABES Institute of Technology, Ghaziabad, U.P., India

ABSTRACT

In last few years the non-technical losses (NTL) identification in electric power distribution system is the utmost importance for the distribution companies. But it is difficult to perform the non-technical losses identification task asthere may be similarities between normal and dishonest customers as far as the consumption patterns are concerned. Therefore, it is very important to improve the reliability of the NTL identification algorithm becomes very important. Non-Technical Losses may be categorized as - theft of electricity, illegal connection, fault metering and billing error. To detect the NTL using classification algorithm,data mining techniques are usually used by implementing an intelligent computational tool to identify the NTL and to select its most close feature byanalysing information from the database of the consumers profiles. This paper reviewsthe k-means algorithm as classification techniques to detect the NTL and also describes its merits as well as drawbacks for the proposed objectives.

The k-meansalgorithm,one of the popular clustering techniques, is used to reduce the average squared distance as far as possible between different points in the same cluster. Here in practice the accuracy is not guaranteed, but its simplicity and the execution speed are very appealing. An optimal clustering algorithm can be achieved if k-means is enriched with a simple and randomized seeding technique. Initial experiments show that it can largely improve both speed and accuracy of k-means. A survey recently made on data mining techniques concludes that the k-means is the most popular clustering algorithm used in scientific and industrial applications [1].

KEYWORDS

non-technical losses, data mining, k-means, optimum-path forest

1. INTRODUCTION

The Non-Technical Losses (NTL), which is a serious concern for the electricity distribution companies, refers to the energy consumed illegally without paying the bill. It creates an unpredictable mismatch between pre-decided consumption to the distribution company and the actual consumption of the customers. This leads to major economic losses for the distribution company and it may be up to 40% of the total electricity distributed in countries such as India, Brazil, Malaysia[2], [3]. The geographic location, economic activity, customer's consumption pattern, the contact demand etc are the features that are used for detection of non-technical losses. However, distribution companies have a detailed database with lots of information about customers profile. Currently, there are no specialized techniques to treat the additional information which can be identified in the existing

database of the company. Usually, they are restricted to use consumption data and limited information related to the customer.

At present, the method of detecting NTL is to conduct physical inspections at the customer premises based on the assumption that there may be chances of disputes.The inspection results are then used in the learning of algorithms to improve the predictions. The number of false positives should be reduced and hence accurate predictions are very important; as physical inspections are expensive due to requirement of technical manpower.

In recent years the artificial intelligence techniques are implemented to identify NTL automatically in smart grids. Despite the extensive use of machine learning techniques for NTL identification, the main challengehere is the data processing in large datasets. There are various proposed methods for the problem of overlapping samples in an effective manner.The applications where retraining at each and every time step is required, the computational burden for training may be restricted.

Classification and Clustering techniques are considered as the fundamental tasks in Data Mining.Classification is mainly used for supervised learning method and forunsupervised learning method clustering is used. In unsupervised learning method clustering is usedto identify a new set of categories, thenew groups are of interest in themselves, and their assessment is intrinsic. But inclassification tasks, however, an important part of the assessment is extrinsic,since the groups must reflect some reference set of classes.

In machine learning and computational geometry,Clustering is considered to be one of the classic problems. In the k-meansmethod, an arbitrary integer ‘k’ and a predefined set of ‘n’ data points to its closest centroid are considered. The objectiveof this method is to choose ‘k’centresto minimizethe sum of the squared distances ϕ, between each point and its closest centre. Lloyd’s algorithm, usually consideredasthe simplified formof k-means, also begins with ‘k’ arbitrary centres, selected randomly from the uniform data points. Each point is designated to the nearest centre, and each centre is then re-evaluated as the centre of mass of all points assigned to it. These two steps i.e. assignment and centre calculation are repeated till the process becomes stabile. It can be seen that the total error is continuously decreasing, which ensure that no clustering is repeated during the execution of the algorithm. As in this scenario there are k^nnumber of possible clustering options, so the iteration will always be terminated. Practically it seems to be faster than any other algorithmas very few iterationsare required here [4].

2. NON-TECHNICAL LOSSES

The Non-Technical Losses (NTL), which is a serious concern for the electricity distribution companies, refers to the energy consumed illegally without paying the bill. It creates an unpredictable mismatch between pre-decided consumption to the distribution company and the actual consumption of the customers. There are various possible causes of non-technical losses, such as different fraudulent types of customers etc, hence NTL detection is a challenging task. As per the machine learning is concerned, one of the main problems is the data imbalance because there are more legal customers as compared to fraudulent types of customers in the distribution systems. This problem has not been addressed and reported in the literature adequately,so various prediction models have been consideredfor different proportions of NTL in the data and reviewedcomparative performance measures for a decisive assessment.

A generalsurveymethod is provided in [5] for NTL detection. Mainly two approaches of fraud detection are discussed here, these are i) ‘expert systems that represent domain knowledge in order to make decisions typically using hand-crafted rules’ and ii) ‘data mining or machine learning techniques that employ statistics to learn patterns from sample data in order to make decisions for future unseen data’. Both approaches have their

validation and neither can be considered as better nor worse than other in artificial intelligence [6].

It may be noted that mostly supervised learning methods are common for NTL detection. Anomaly detection, which may be considered as a superclass of NTL detection, is more challenging for supervised analysisdue tofollowing reasons, i) imbalanced classes are created when anomaly data sets contain a very small number of positive examples and large number of negative examples, ii) it is used for many different kinds of anomalies as it is hard for any algorithm to learn from just a few positive examples what the anomalies might look like and iii) there may be also future anomalies which may look completely different to any of the anomalous examples learned so far [7]. On the other hand, supervised learning method works betterbecause of i) availability of large numbers of both positive and negative examples, ii) the algorithm gets a sense of what positive examples might look like, when there are enough positive examples and iii) future positive examples are likely to be similar to the ones in the training set.

3. K-MEANS ALGORITHM

The k-means algorithm is a simple iterative method to split a given dataset into a user-specified number of clusters, k. In different period of time various scientists have developed the k-means algorithm, e.g., Lloyd (1957), Forgey (1965) and McQueen (1967) [8]. The brief history and the types of k-means algorithm has been discussed in [9]. Gray and Neuhoff [10] presentedthe historical background where it is placed in the larger context of hill-climbing algorithms.

The k-means algorithm, one of the popular clustering techniques, is used to reduce the average squared distance as far as possible between different points in the same cluster.In this technique the number of clusters are considered as an input, andit groups data trying to separate samples in groups with similarvariance in order to minimize some criterion. This algorithm partitions the data into $k$clusters ($C_1, C_2, \ldots, C_k$), represented by their centres or means. The mean of all the instances belongs to a particular cluster is considered as the centre of individual clusters. This algorithm scaleswell to large number of samples, and it has been widely used inmany different application areas [11,12].The k-means algorithm iterates between two steps,mentioned bellow, till the convergence is achieved:

Step 1: It is the ‘Assignment’ stepwhere each data point is assigned to the most similarly generated model parameters to the nearest centroid. This step is usually used for partitioning of the data.

Step 2: In the ‘Relocation’ step,each cluster model is relocated to the centre(mean) of all data points assigned to it. It should be noted that if a possibility measure is achieved for the data points, then it is expected that the relocation to be done for the data partitions. If the assignments do not change further the algorithm is supposed to be converged.

It works as follows.

1. Arbitrarily choose k initial centres $C = \{c_1, \ldots, c_k\}$.

2. For each $i \in \{1, \ldots, k\}$, set the cluster C_i to be the set of points in χ that are closer to c_i than they are to c_j for all $j \neq i$.

3. For each $i \in \{1, \ldots, k\}$, set c_i to be the centre of mass of all points in C_i: $c_i = \frac{1}{|C_i|}\sum_{x \in C_i} x$

4. Repeat steps 2 and 3 till C becomes constant.

It is a general practice to choose the initial centres uniformly at random from χ. For Step 2, ties may be broken arbitrarily, as long as the method is compatible.

In steps 2 and 3, the squared distances $\phi$is definitely decreased,as a result the algorithm can be improvedfor an arbitrary local clustering, chosen randomly. To see that Step 3 does in fact decreases the squared distances ϕ, a standard result from linear algebra can be recalled.

To start k-means algorithm initially a set of cluster centres chosen at random or according to some analytical procedure. In each iteration, each occurrence is assigned to its nearest cluster centre according to the Euclidean distance between the two. Then the cluster centres are reassigned. The centre of each cluster is calculated as the mean of all the occurrences belonging to that cluster:

$$\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} X_q \qquad (1)$$

where $N_k$is the number of occurrences belonging to cluster *k* and $\mu_k$is mean of the cluster *k*.

These selected centres implicitly define a clustering – for each centre, randomly selected cluster are supposed to be the set of data points that are nearer to that centre than to any other. As noted above, finding an exact solution to the k-means problem is difficult but to achieve a number of convergence conditions are possible. For example, it may terminate the search option when the partitioning error is not reduced by the relocation of the centres. Hence the present partition is considered to be optimal within the limited area. Other stopping criteria can be used also such as exceeding a predefined number of iterations.The main aim of the k-means algorithm isto choosethe centroids, so that the within-cluster sum of squared criterion [10] can be minimized, given by:

$$\sum_{i=0}^{N} \min_{\mu_j \in C_j} \left(\left\| x_i - \mu_j \right\|^2 \right), \forall\ j = 1, 2, \ldots, k, \qquad (2)$$

being $x_i$sample from the training data.

Unfortunately, at the price of accuracy, the simplicity and the speed of the k-means algorithm is achieved. There are many practical examples where the algorithm generates arbitrarily bad clustering. It never confirms the adversarial placement of the initially chosen centre. Even with the standard randomized seeding technique,with high probability, the difference may be unlimited [1].

The algorithm is very sensitive to the initial selection of cluster centres, which may make the difference between global and local minimum. Like single link algorithms, the k-means algorithm is not versatile because it is a partitioning algorithm which works only on isotropic clusters of data set.It is most acceptable because of:

1. Its time complexity *O*(*m***k***l*), where *m* is the total number of instances; *k*is the number of clusters to be considered and *l* is the number of iterations required for convergence. Initially*k* and *l* are made fixed, so thealgorithm has linear time complexity in the size of the data set.

2. Its space complexity is *O*(*k*+*m*), requires more space to store thedata matrix. The data matrix can be stored in a secondary memoryand when required each pattern can be accessed separately. Here processing time requirement is very high because of iterative nature of the algorithm.

3. It is order-independent. Whatever may be the order of patterns, presented to the algorithm, it always generates the same partition for a given initial set of clusterscentres.However, the performance of the k-means algorithm is highly dependent on the initial seed selection so that it can produce only hyper-spherical clusters at the best.

4. EVALUATION ANALYSIS

For evaluation two distinct processesis conducted i.e.i)related to the unsupervised NTL identification and ii) related to the anomaly detection task. The available information can

beused to design the datasets, but such knowledge is not sufficientduring the learning process. For anomaly detectiontask, a new dataset composed of legal consumer profiles to bebuilt. These data are considered as the "normal" samples, and just afterthe unsupervised learning process executionthe sample,which does not fitinto the learned model, is then considered as an "anomaly".

Considering both processes, Papa et al. [13] proposes a classification accuracy that considers unbalanced datasets, which is a common problem in the context of NTLdetection.Such classification accuracy strongly penalizes errors onsmall classes, being adequate to the context of this work, since thereare much more samples from regular consumers than irregularones. The accuracy is measured by taking into account the classesmay have different sizes in a dataset $\mathcal{Z}$. Let us define:

$$e_{i,1} = \frac{FP_i}{|\mathcal{Z}| - |\mathcal{Z}^i|} \quad (3)$$

and

$$e_{i,2} = \frac{FN_i}{|\mathcal{Z}^i|} \ , i = 1, 2, \ldots, K \quad (4)$$

where $\mathcal{Z}$stands for the number of classes, $|\mathcal{Z}^i|$ concerns with thenumber of samples in $\mathcal{Z}$ that come from class i, and $FP_i$and FN_istandfor the false positives and false negatives for class i, respectively.That is,FP_i is the number of samples from other classes that wereclassified as being from the class i in $\mathcal{Z}$, and $FN_i$is the number ofsamples from the class i that were incorrectly classified as beingfrom other classes in $\mathcal{Z}$. The error terms $e_{i,1}$,and $e_{i,2}$are then usedto define the total error from class i:

$$E_i = e_{i,1} + e_{i,2} \quad (5)$$

Finally, the accuracy A_{cc} is then defined as follows:

$$A_{cc} = 1 - \frac{\sum_{i=1}^{K} E_i}{2K} \quad (6)$$

Apart from the previous accuracy measure, the F-measure, which is a recognized measure, is also employed [14] thatconsiders both the precision and recall information.

Numerical characteristics is the foundation of the k-means algorithm. Noisy data affect this algorithm severely and the number of clusters should be available in advance, which is not insignificantfor unavailability of prior knowledge. Haung (1998) presented the *K*-prototypes algorithm, based on the k-means algorithm, where numeric data limitations are removed but the efficiency isretained. The algorithm clusters objects similar to the k-means algorithm with numeric and categorical attributes. The square Euclidean distance is the similarity measure on numeric attributes and the number of mismatches between objects and the cluster prototypes is the similarity measure on the categorical attributes.

5. GENERALIZATIONS AND CONNECTIONS

It has already been discussed that the k-means is directlyassociated to fitting a mixture of k isotropic Gaussians to the data. Further, fitting the data with a mixture of k components from the exponential family of distributions is directly related to the 'generalization' of the distance measure to all divergences. Another important 'generalization' is to view the "means" as the probabilistic models instead of points to its closest centroid. Next in the 'assignment' step, each data point is assigned to the most similarly generated model parameters. The 'relocation' step is used to update the model parameters so that it can fit the assigned datasets with higher accuracy. More complex data can be provided in this model-based k-means algorithm.

In the implicit high-dimensional space, the boundaries between clusters are usually linear which may be converted to non-linear when projected back to the original space.So, it deals

with more complex clusters by allowing kernel k-means. Dhillon et al. [15] has shown the close connection between kernel k-means and spectral clustering.

The k-means algorithm, may have some drawbacks, still is considered asone of the popular methods used for partitional clustering techniques for analysis of unsupervised data. This algorithm is simple, easily understandable and reasonably scalable. Also, it can be easily modified to deal with streaming data. The execution speed of k-means is required to increase with significant efforts when very large data sets are required to deal.For any algorithm there should be amendments for continuous improvement and generalization to ensureits relevance and also to improve the effectiveness.

6. LIMITATIONS

Although the k-means algorithm is the simplest and most commonly used algorithm but it also has some drawbacks apart from being sensitive to initialization. By using k-means, data is basically fitted with identical, isotropic covariance matrices ($\Sigma = \sigma^2 I$) with a mixture of k Gaussians, when the soft assignments of data points to mixture components become difficult to allocate individual data point only to the most similar component. So, if the data is not well described, it will be collapsed;for example, non-covex shaped clusters in the data. By rescaling the data before clustering this problemcan be reduced.Further, a different distance measure may bemore relevant for the dataset. For example, theKL-divergence represents two distinct probability handling, this is because of measuring the distance between two data points for information related clustering. It has been recently shown that the primary properties of k-means i.e. guaranteed convergence, linear separation boundaries and scalability, are retained if the distance is measured by selecting any cluster of a very large class of divergences called Bregman divergences during the assignment step and makes no other changes [16]. As a result, for a much larger class of datasets,k-means becomes more effective till an appropriate divergence is used.

The k-meansalgorithm is paired with another algorithm to describe the non-convex clusters. Initially unsupervised data are clustered into a huge number of groups by using k-means algorithm. Now these groups are clustered into larger clusters using single contact hierarchical clustering, which is capable of detecting complex shapes. Using this approach, the solution becomes less sensitive for initialization.Since the hierarchical method yields results with multiple resolutions, itneed not to pre-specify ‘k’further.

When thenumber of clusters equals the number of distinct data-points, the cost of optimal solution process decreases with increase in the number of ‘k’ till it hits zero. As a result, itbecomes difficult to compare solutions with distinct numbers of clusters directly andit is also challenging to find the optimum value of ‘k’. The k-means algorithm has to be run with different values of ‘k’,if the desired value of ‘k’ is not known in advance and then optimum result has been selected by using a suitable criterion. Usually, in k-means methodology an arbitrary complex term, increases with ‘k’,is added to the original cost function (Eq. 2) and then the cost is supposed to be minimized by selecting appropriate value of ‘k’.On the other hand, the number of clusters can be increasedgradually along with a suitable termination criterion. This is achieved in bisecting k-meansalgorithm [17] by putting all the data into one cluster first, and then the least compact cluster are split into two parts by sung k-means. The LBG algorithm [10], a popular algorithm for large data sets mainly used for vector quantization,makes the number of clusters as double until a suitable solution is obtained. Both these approaches thus mitigate the need to know ‘k’in advance.

Due to the presence of anomalies, this algorithm becomes sensitive, becausethe “mean” is not an in good shapestatistic for the analysis. It may be helpful to opt apre-processing step to remove anomalies. The resultsshould further be post-processed to eliminate small clusters or tomerge close clusters into a large cluster. Ball and Hall’s ISODATA algorithm from 1967 effectively used both pre-processing and post-processing on k-means [18].

7. OPTIMUM-PATH FOREST

Recently, Optimum-Path Forest (OPF), which is a different framework of graph-based classifiers, is introduced. It can reduce the pattern recognition problem in the feature space induced by that graph [13]. This kind of classifier does not explain the classification task as a hyperplane optimization problem, but rather as a combinatorial optimum-path computation based on specific key samples (prototypes) to the remaining nodes.

Such classifiers never interrupt the classification tasks, i.e. hyper planes optimization issues whereas, the combinatorial optimum-path computation is available for the remaining nodes. Each node is categories as per the strength of the connected prototypes, defined as the discrete optimal partition for the feature space.

In the proposed method of OPF algorithm, each pattern is supposed to be the root for its optimum-path tree, and each of the node is classified according to the significance of its connection to the pattern, which defines a discrete impact region (optimal partition) of the feature space. Ramos, C.C.O. et al proposed a technique, which is based on the OPF to identify and to protect the unsupervised non-technical losses. It is obvious that the OPF has better probability to identify the electricity theft; also, they are even higher in accuracy when compared to GMM.

OPF-based classifiers have the following advantages, i) they run the training phase faster so real-time applications for fraud detection in electrical systems are possible, ii) they do not assume any special shape or separability of the feature space, and iii) they are free of parameters [19].

8. CONCLUSION AND FUTURE WORK

In this work, k-meansalgorithm with the optimalclustering has been presented. Moreover, the seeding technique is fasteras well as simple whichmakes it attractive in practice. Usually, the standard practice is that the k-means algorithm has to run multiple times, and then keep only the bestclustering found, it is likely to achieve a constant approximation at least once.But it does not confirm whetherthis algorithm achieves better results if it isallowed a greater number of trials or a similar result can be achieved for asmaller number of trials.

In summary, for unsupervised data sets the k-means algorithm can be applied to find the NTLs; other available algorithms can be applied on small data sets. It is difficult to get suitable learning/control parameters for ANNs, GAs, TS, and SA and for large data set the executiontime is also very high. But it has already been discussed that the k-means algorithm converges to an optimal local solution. This behaviour is linked with the initial seed selection in the k-means algorithm. So, the k-means algorithm would work better for large data sets also if an initial partition is obtained instantly by using any other process.

REFERENCES

[1] Pavel Berkhin,“Survey of clustering data mining techniques”, Technical report, Accrue Software, San Jose, CA, 2002.

[2] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni and R. C. Green, “High Performance Computing for Detection of Electricity Theft”, International Journal of Electrical Power & Energy Systems, vol. 47, issue 1, pp. 21-30, May 2013.

[3] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed and F. Nagi, “Improving SVM-Based Nontechnical Loss Detection in Power Utility Using the Fuzzy Inference System”, IEEE Transactions on Power Delivery, vol. 26, issue 2, pp. 1284-1285, April 2011.

[4] Arthur D. and Vassilvitskii S.,“k-means++: the advantages of careful seeding”, In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1025, 2007.

[5] Y. Kou, C.T. Lu, S. Sirwongwattana and Y.P. Huang, "Survey of fraud detection techniques", IEEE International Conference on Networking, Sensing and Control, vol. 2, pp. 749-754, 2004.

[6] D. Gorgevik, D. Cakmakov and V. Radevski, "Handwritten digit recognition using statistical and rule-based decision fusion", 11th Mediterranean Electrotechnical Conference (MELECON), pp.131-135, 2002.

[7] A. Ng, "Machine Learning", Coursera, 2014.

[8] Lloyd SP (1957) Least squares quantization in PCM. Unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical Statistics Meeting Atlantic City, NJ, September 1957. Also, IEEE Trans Inform Theory (Special Issue on Quantization), vol IT-28, pp 129–137, March 1982.

[9] Jain A. K., Dubes R. C., "Algorithms for clustering data", Prentice-Hall, Englewood Cliffs, 1988.

[10] Gray R. M., Neuhoff D. L., "Quantization", IEEE Trans Inform Theory 44(6):2325–2384, 1998.

[11] C. Manning, P. Raghavan, H. Schütze, "Introduction to Information Retrieval", Cambridge University Press, New York, NY, USA, 2008.

[12] Scikit-Learn Developers, User Guide, 2016, Available at http://scikit-learn.org/ dev/user guide.html.

[13] J. Papa, A. Falcao, C. Suzuki, "Supervised pattern classification based on optimum-path forest", Int. J. Imaging Syst. Technol. Vol. 19, pp 120–131, 2009.

[14] D.M.W. Powers, "Evaluation: from precision, recall and F-measure to roc, informedness, markedness and correlation", Int. J. Mach. Learn. Technol. Vol. 2, pp 37–63., 2011.

[15] Dhillon IS, Guan Y, Kulis B, "Kernel k-means: spectral clustering and normalized cuts", KDD 2004, pp 551–556, 2004.

[16] Banerjee A, Merugu S, Dhillon I, Ghosh J, "Clustering with Bregman divergences" J Mach Learn Res, vol. 6, pp 1705–1749, 2005.

[17] Steinbach M, Karypis G, Kumar V, "A comparison of document clustering techniques", In: Proceedings of the KDD Workshop on Text Mining, 2000.

[18] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, B. Liu, M Steinbach, D. J. Hand, D. Steinberg, "Top 10 algorithms in data mining", Springer, Vol. 14, Issue 1, pp. 1-37, 2008.

[19] C.C.O. Ramos, A.N. de Sousa, J.P. Papa, A.X. Falcao, "Learning to Identify Non-Technical Losses with Optimum-Path Forest", In: Proceedings of the IWSSIP-2010-17th international conference on systems, signals and image processing, pp. 1–4, 2010.