



UNIVERSITÉ
JEAN MONNET
SPÉCIALITÉ WEB
INTELLIGENCE

TSINGHUQ
UNIVERSITY
DEPARTMENT OF
ELECTRONIC ENGINEERING

STAGE EN ENTREPRISE: STAGE DE FIN D'ÉTUDE

Rapport de Stage Année 2013-2014

Fouille de Donnée dans la domaine de
télécommunication

Auteur :
Wenyi WANG

Tuteur de stage en entreprise :
Vice directeur de labo NGN :
yongfeng HUANG
Tuteur de l'université :
Amaury HABRARD

De 20 Février 2014 à 20 Juillet 2014

Table des matières

Remerciements	3
1 Résumé	1
2 Introduction	1
2.1 Introduction du CMCC	1
2.2 La crise du CMCC	2
2.2.1 La demande de mise à niveau du réseau	2
2.2.2 Les changements des moyennes de revenu	4
2.3 L'optimisation du réseau	5
2.4 Introduction du laboratoire	6
2.5 Objectif du projet	6
3 Introduction de l'industrie de la télécommunication	7
3.1 L'évolution des normes de téléphonie mobile	7
3.1.1 La première génération	8
3.1.2 La deuxième génération	8
3.1.3 La troisième génération	9
3.1.4 La quatrième génération	9
3.2 Le réseau LTE	10
3.2.1 La structure du réseau LTE	11
4 Les solutions existantes	13
4.1 la Technique KQI	13
4.2 la Technique QoE	14
4.3 la Technique qui étudie les comportements de l'utilisateur	16
5 Le présentation de notre solution	16
5.1 Clustering	17
5.1.1 La distance	18

5.1.2	La validité de clustering	19
5.1.3	La silhouette Coefficient	20
5.2	Règles d'association	21
5.3	K-Means et l'arbre couvrant de poids minimal	22
5.3.1	L'algorithme KmMST	23
6	La mise en œuvre	24
6.1	Le logiciel utilisé	24
6.2	Introduction des données	25
6.3	Pré-traitement de données	26
6.4	Les caractéristiques de données	29
6.5	Le K-Means et Le Règle d'association	30
6.5.1	La valeur optimale de k	30
6.5.2	Le clustering	31
6.5.3	La Règle d'association	33
6.6	Le KmMST	36
6.6.1	L'étape de l'algorithme KmMST	36
	Conclusion	38
	Références	39

Remerciements

Tout d'abord, je tiens à remercier Amaury Habrard et tous les enseignants de la Spécialité Web Intelligence de l'Université Jean Monnet, aussi les enseignants de Télécom Saint-Etienne et L'école nationale supérieure de Saint-Etienne, qui m'a aidé lors de ces deux années de étude.

Je remercie également M.Yongfeng HUANG pour avoir accepté diriger cette stage, il m'a beaucoup conseillé, et les discussions que l'on a pu avoir se sont toujours révélées très intéressantes et instructives.

Je souhaite également adresser mes remerciement à Zheng YANG, Lindong WEI et xian WU ainsi que tout les membres du laboratoire de Next generation Network(**NGN**) pour m'avoir soutenu, encouragé et conseillé tout au long de ce stage.

Je tiens à montrer tout ma gratitude envers toutes les personnes qui ont pu m'aider, m'encourager, me soutenir, me remotiver pendant ces années de travail.

1 Résumé

Pendant ces quatre mois de stage, notre groupe de recherche travaille avec les employés de CMCC (China Mobile Communications Corporation) . L'objectif du sujet est : utiliser les techniques de Fouille de données, étudier les données fournies par le CMCC, et trouver les relations entre les données et les défauts du système 4G. Nous avons fait plusieurs tentatives pour trouver les résultats, et on a utilisé de différents logiciels, j'ai utilisé le R, et mon collègue utilisé Matlab, nous avons utilisé plusieurs algorithmes (Clustering, PCA, Association rules, Ajustement). Mais à la fin, nous avons trouvé qu'à cause des défaut dans la système d'acquisition, les données ne sont pas correct, et nous ne pouvons pas trouver le résultat comme prévu. Mais les recherches que nous avons fait peuvent faire savoir comment utiliser les technique de fouille de données dans le domaine de télécommunication.

2 Introduction

Le 3 avril 1973, M. Mation COOPER le directeur général de la division communication de Motorola, a effectué un appel téléphonique à Joel ENGEL, son rival et néanmoins confrère chez Belle Labs. c'est le premier appel téléphonique en extérieur, L'idée du téléphone portable devient une réalité.

Depuis ce jour, la technique développe très rapidement. Pendant les 20 dernières années, il y a déjà quatre générations des standards pour le téléphonie mobile, non seulement nous pouvons appeler les autres, les nouvelles technologies et les Smart-phones nous permettons aussi d'envoyer les messages, de surfer sur l'Internet, d'utiliser le service RTSP(Real Time Streaming Protocol), et le service VoIP (Voice over Internet Protocole),etc.. Les services de communication téléphonique sont devenus un outil très important dans notre vie.

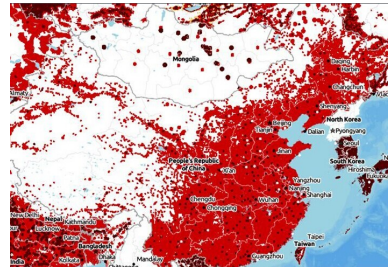
2.1 Introduction du CMCC

Fondé en 3 septembre 1997, après le regroupement d'opérateurs des télécommunications en 2008, CHINA MOBILE COMMUNICATIONS CORPORATION (CMCC)[1\(a\)](#) est devenu un des trois opérateurs des télécommunications en Chine (deux autres sont China Unicom Co., Ltd. et China Telecom). Après plusieurs années de développement, il a construit un du plus plus grand réseau de communications mobiles dans le monde, possède la plus grande base

d'utilisateurs dans le monde1(b). En 2013, le CMCC a 767 million utilisateurs, 630,2 billion ¥ de revenu, 121,7 billions ¥ de revenus net, et un effectif de 197,030 personnes.



(a) Logo de China Mobile



(b) Réseau télécommunication

FIGURE 1 – CMCC

2.2 La crise du CMCC

2.2.1 La demande de mise à niveau du réseau

Mais en même temps, le taux de croissance des nouveaux utilisateurs décline de 22,5 % (2006) à moins de 5% en 2013 2. Et pour les trois premières mois, bien que l'entreprise soit une fois considéré comme la plus rentable de Chine, le taux de croissance des revenu net n'est que 0,3%.

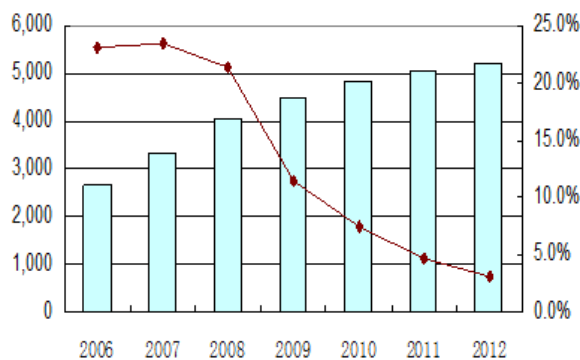
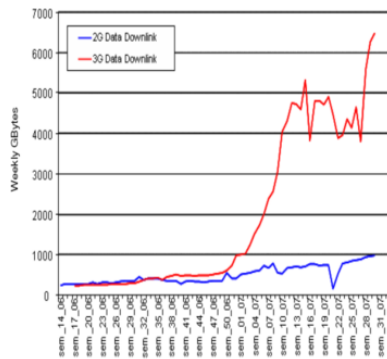


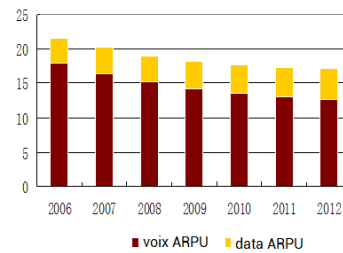
FIGURE 2 – le taux de croissance est décliner

Opérateur des télécommunications Vodafone a fait une étude après qu'il a déployé un réseau 3G(the third generation of mobile phone mobile communication technology standards). Comme le réseau 3G permet des débits (de 2 à 42 Mb/s définis par la dernière génération des réseaux) qui sont bien plus rapides

que la génération précédente, par exemple le GSM. Les utilisateurs utilisent bien plus souvent le service internet^{3(a)}. Comme ils utilisent plus du service internet, le data ARPU (Average Revenue Per User) augmente, mais le voix ARPU décline plus rapidement que l'augmentation de data ARPU^{3(b)}.



(a) Downlink Data Traffic in 2G/3G Network



(b) étude de Vodafone

FIGURE 3 – Vodafone

Mais l'étude d'Orange nous montre que si nous pouvons fournir des nouvelles technologies qui ont plus haute débit, les utilisateurs utiliseront plus souvent le service data. ⁴

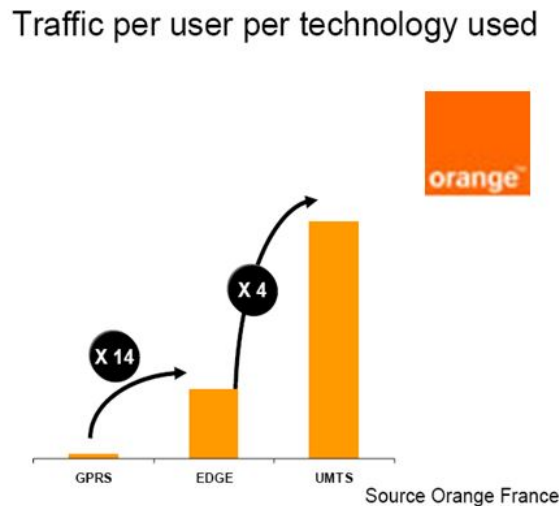


FIGURE 4 – trafic par personne

Des études nous montrent que de nouvelles technologies (comme LTE)

peuvent diminuer le prix de revient qui peut assurer le profit de l'opérateur. Mais déployer les nouveaux matériels coûte cher, en 2009, le CMCC a dépensé 30 milliards ¥ pour construire les stations de réseau 3G, et en 2014, le CMCC a construit 1,5 million stations, à la fin de cette année, il y aura 1,8 million stations, parmi ces stations, il y aura 500 mille stations TD-LTE. En ajoutant des équipements 4G, il peut être mis à niveau une station de 3G à 4G. Donc déployer le réseau 4G n'est pas trop cher, selon l'expérience précédente (de 2G à 3G), les utilisateurs vont utiliser plus le service internet, qui peut assurer le profit de l'entreprise.

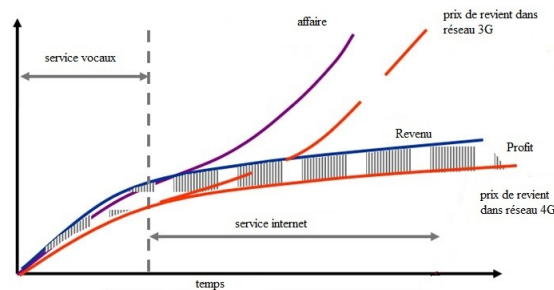


FIGURE 5 – 4G est plus rentable

2.2.2 Les changements des moyennes de revenu

Avant la popularité des téléphones intelligents, et avant la popularité du réseau 3G, beaucoup des utilisateurs du CMCC utilisent les services message pour recevoir des informations comme les nouvelles, les météo, etc. Ces services étaient un moyen de revenu très important pour le CMCC. Mais maintenant, avec la popularité du smartphone et l'évolution du réseau, les gens peuvent trouver toutes les informations sur l'internet. En la vacance du nouvel an chinois 2013, 31,3 milliards SMS ont été envoyés, comparé à cette année, le nombre est 18,2 milliards SMS, il a réduit 42%. Ces informations sont très inquiétantes pour le CMCC.

En conclusion, CMCC doit mettre à jour son réseau de télécommunication, mais la mise à niveau du réseau entraîne une réduction des revenus des autres services. Alors CMCC veut apprendre les entreprises comme Tencent et trouver des moyens pour augmenter le revenu.

2.3 L'optimisation du réseau

En plus de l'évolution des technologies, un grand enjeu pour les opérateurs est : l'optimisation du réseau télécommunication. Le réseau de communication mobile est très dynamique, la répartition de la densité du trafic est inégale, fréquence très limitée, etc. La configuration du réseau était toujours sous-optimale, et la perception de l'utilisateur n'est pas très bien. Donc tous les opérateurs doivent toujours reconfigurer/optimiser/maintenir les paramètres du réseau.

Les opérateurs peuvent percevoir les données sur Internet, et utilisent ces informations pour trouver les défauts du système, ce peut aider l'entreprise optimiser le système.

Mais l'optimisation du réseau télécommunication est difficile, parce-que les technologies d'optimisation de réseau concerné à savoir la technologie de commutation, la technologie sans fil, la configuration et commutation de la fréquence, la signalisation système, l'analyse de trafic, etc. sont des travaux difficiles, qui exigent une meilleure aptitude des employés.

Actuellement, l'optimisation du réseau dépend principalement de l'expérience du personnel. Mais des fois, les expériences ne sont pas correctes. Par exemple, si l'entreprise a besoin de savoir le congestionné d'une station, il faut envoyer les employés avec des équipements pendant les périodes de pointe, mais on ne sait pas si les résultats sont corrects [6](#). En outre, souvent un seul type de donnée est utilisé pour l'analyse et la comparaison pour optimiser les réseaux, il faut mieux de trouver une solution d'optimisation basée sur toutes les données liées au réseau (telles que les données statistiques de trafic, les données d'essai, etc). Et en raison de l'énorme quantité de données, c'est difficile de traiter en temps opportun. Il est évident que cette méthode est défectueuse. Les défauts du système provoquent la non satisfaction des utilisateurs, ce qui a conduit les défauts à se multiplier.

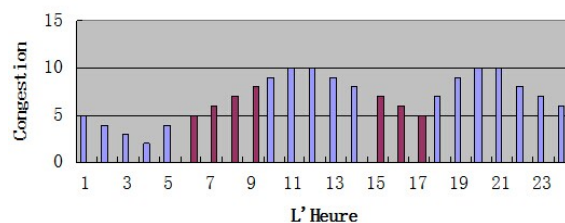


FIGURE 6 – Mesure la congestionné d'un station

Face à des problèmes complexes, les grandes entreprises commencent à

utiliser les techniques de Fouille de données. Cette technique peut aider l'entreprise à prendre faire les décisions plus vite et plus précises.

De ce faire, en Juillet 2013, le CMCC a lancé ce projet avec quatre laboratoires dans trois universités, ils sont [Tsinghua University](#), [Shandong University](#) et [University of Electronic Science and Technology of China](#). Le projet inclut trois partiel : Fouille de données, gestion du Cloud plateforme et modélisation de l'information dans le système.

2.4 Introduction du laboratoire

De 20 Avril 2014 à 20 Juillet 2014, je fait mon stage chez [laboratoire of Next Generation Network Technology & Application \(NGN\)](#) 7. C'est d'un subordonné de [Research Institute of Network And Human-Machine Speech Communication](#), Département Ingénierie électronique, Tsinghua University. Le laboratoire se trouve dans le ROHM bâtiment.



FIGURE 7 – Logo NGN

Les principaux axes de recherche sont Théorie des réseaux, Architecture de l'Internet, Traitement de l'information Internet, recherche dans le domaine de la sécurité Internet, Sentiment analyse, Information hiding, etc.

Mon tuteur professionnel est [M. Yongfeng HUANG](#), vice-directeur du laboratoire NGN. Dans le laboratoire, il y a cinq groupes, chaque groupe a un docteur et son sujet. dans notre groupe, il y a quatre personnes, un étudiant de doctorat de la première année, un étudiant de M1, et une étudiante de Licence de la troisième année et moi. On utilise R et Rstudio, et Hadoop aussi.

2.5 Objectif du projet

Dans cet article, nous allons d'abord présenter le réseau de communication mobile, ensuite décrire l'état de l'optimisation du réseau et les techniques pour l'optimisé. pour enfin je présenter notre solution et le conclusion de ce stage.

3 Introduction de l'industrie de la télécommunication

3.1 L'évolution des normes de téléphonie mobile

Depuis 1984, il y a déjà plusieurs standards ont été utilisés par les opérateurs dans le monde entier. Voici un tableau de différents standards de mobile en Europe et les paramétrés 1.

Génération	Acronyme	Description	Débit
1G	Radiocom 2000	Échanges de type voix uniquement	analogique
2G	GSM	Échanges de type voix uniquement	9,05 kbps
2,5G	GPRS	Échange de données sauf voix	171,2 kbps / 50 kbps / 17,9 kbps
3G	UMTS	Voix + données	144 kbps rurale, 384 kbps urbaine, 1,9 Mbps point fixe / -
3.5G ou 3G+ ou H	HSPA	Évolution de l'UMTS	14,4 Mbps / 3,6 Mbps / -
4G	LTE	Long Term Evolution (Données)	150 Mbps / 40 Mbps / -
4G	LTE-Advanced	Long Term Evolution Advanced (Données+voix)	1 Gbps à l'arrêt, 100 Mbps en mouvement / - / -

TABLE 1 – Les différentes générations de téléphonie mobile en Europe

3.1.1 La première génération

En télécommunication, 1G est la première génération des standards pour la téléphonie mobile, Il s'agit de la première apparition du réseaux de téléphonie mobile. 1G consiste aux réseaux analogiques qui peuvent échanges de type voix uniquement.

3.1.2 La deuxième génération

2G, la technologie de téléphonie sans fil de deuxième génération, la différence entre les réseaux 1G et 2G est : les signaux radio sur les réseaux 1G sont analogiques, et ceux de 2G sont numériques.

Systèmes 2G ont été significativement plus efficaces du spectre permettant de bien plus grands taux de pénétration du téléphone mobile, en plus les données vocales numériques peuvent être compressées et multiplexées beaucoup plus efficacement que les codages de la voix analogique grâce à l'utilisation de codecs différents, ce qui permet plus d'appels à transmettre dans la même quantité de bande passante radio. Et 2G introduit pour la première fois le service de données pour mobile. La Technologie 2G permettent les divers réseaux de téléphonie mobile d'utiliser des services tels que le SMS et MMS. Tous les messages de texte envoyés au delà de 2G sont chiffrés numériquement, ce qui permet le transfert de données de telle sorte que seul le destinataire peut recevoir et lire.

Réseaux 2G ont été construits principalement pour le service téléphonique et de transmission de données lente (défini dans les documents de spécifications IMT-2000).

Réseaux 2,5G, qu'on les qualifie souvent de General packet Radio Service ou GPRS, est une norme pour la téléphonie mobile dérivée du GSM et complémentaire de celui-ci, permettant un débit de données plus élevé. Le 2,5 indique que c'est une technologie à mi-chemin entre le GSM (deuxième génération) et l'UMTS (troisième génération). Le GPRS est une extension du protocole GSM : il ajoute par rapport à ce dernier la transmission par paquets. Cette méthode est plus adaptée à la transmission de données. En effet, les ressources ne sont allouées que lorsque des données sont échangées, contrairement au mode « circuit » en GSM où un circuit est établi – et les ressources associées – pour toute la durée de la communication. Le GPRS a ensuite évolué au début des années 2000 vers la norme EDGE également optimisée pour transférer des données et qui utilise les mêmes antennes et les mêmes fréquences radio.

3.1.3 La troisième génération

La troisième génération (3G) des normes de téléphonie mobile. Elle est représentée principalement par W-CDMA, CDMA2000, TD-SCDMA et WiMAX. Elle permet des débits de 2 à 42 Mb/s qui sont bien plus rapides que la génération précédente. Grâce à l'utilisation des règles de classement d'utilisateurs, les bandes de fréquences supérieures rend la capacité du réseau augmenter.

Les différents standard 3G et ses prédécesseurs, utilisent le domaine CS (Circuit Switch) pour le service vocal, et le domaine PS (Packet Switch) s'occupant du service de données [8](#).

3.1.4 La quatrième génération

La quatrième génération des standards pour la téléphonie mobile, succédant à la 2G et la 3G, en théorie, elle permet de transmettre des données à des débits supérieurs à 100 Mb/s.

Une des particularités de la 4G est sa EPC (Evolved Packet Core) basé sur IP, et il n'y a plus de mode commuté (le 'Circuit Switched Domain' qui s'occupe le service vocal dans les standard précédents), ce qui signifie que le service vocal transmis sur l'internet [9](#).

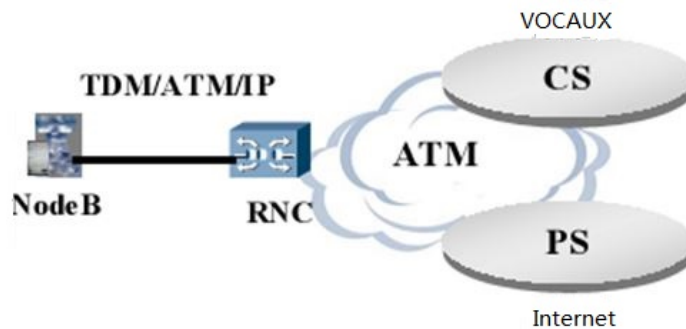


FIGURE 8 – Réseau 3G et ses prédécesseur



FIGURE 9 – Réseau 4G

Les avantages du réseau 4G sont : plus haut débit, mieux utilisation de la bande de fréquence, moins de délai (délai dans le panneau de l'utilisateur est inférieur que 5 ms, délai dans le panneau de commande est inférieur que 100 ms), plus simple structure du réseau, moins de consommation d'énergie terminale.

3.2 Le réseau LTE

Le LTE (Long Term Evolution) est l'évolution la plus récente des normes de CDMA 2000, TD-SCDMA, GSM. La technologie LTE est considérée comme une norme de troisième génération '3.9G', et la 'vraie 4G', appelée LTE-Advanced est reconnue par l'UIT comme une technologie 4G en 2010. LTE a deux branches : LTE-FDD (Frequency-Division Duplex Long Term Evolution) et LTE-TDD, (Time Division Duplex Long Term Evolution) les deux standards sont similaires, la différence entre les deux standards est moins de 15% [10](#). En 2011-2012, les réseaux LTE-TDD sont commercialisés sous l'appellation 4G par le CMCC en Chine.



FIGURE 10 – l'évolution des standard

3.2.1 La structure du réseau LTE

Le réseau 4G contient 3 parties : UE(User Equipment) ;, eNodeB (les stations de base), EPC (Evolved Packet Core). EPC contient MME, S-GW, P-GW et HSS 11 2.

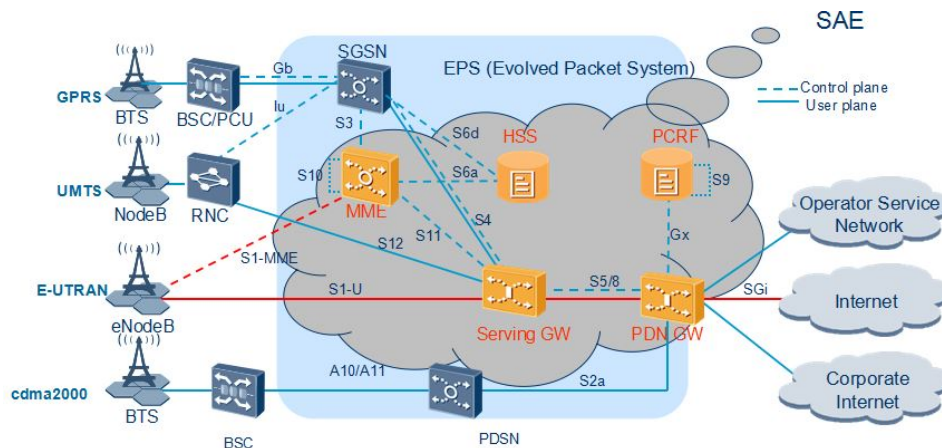


FIGURE 11 – la structure du réseau

Part	Fonction
MME	L'authentification des utilisateurs et la gestion des clés, Cryptage de la couche NAS, Gestion de la liste TA, Sélection P-GW ou S-GW
Service Gateway	Compression d'en-tête IP, Routage de paquets et la transmission, La commutation entre eNB, Facturation des utilisateurs porteur
PDN Gateway	L'allocation des adresses IP de UE, l'accès aux fonction de gestion de réseau externes, Facturation en service
HSS(Home Subscriber Service)	Stockée données de l'utilisateur associées au service
PCRF	Roaming

TABLE 2 – la fonction du chaque partie

Entre deux E-UTRANS, il y a l'interface X2, l'interface S-11 se trouve entre S-GW et MME, E-UTRAN et S-GW échangent les données par l'interface S1-U et il échange les donnée par l'interface S1-AP avec MME, MME et HSS utilisent l'interface S6A, et l'interface S5/8 entre S-GW et P-GW, Gx entre PCRF et P-GW. En mettant des capteurs sur les interfaces, les opérateurs et les fournisseurs d'équipement peuvent collecter les données de signalisation, et utilisent ces informations pour trouver les défauts du système.

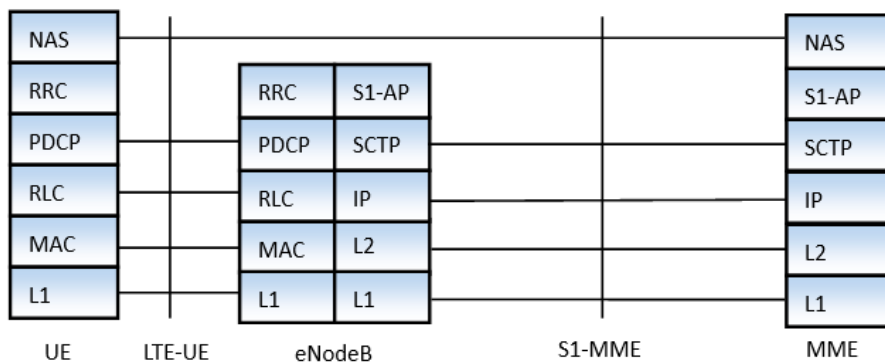


FIGURE 12 – Contrôle plan

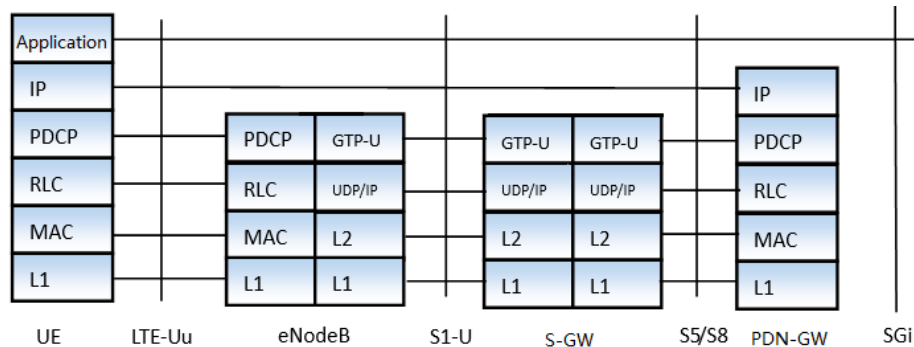


FIGURE 13 – User plan

4 Les solutions existantes

L'optimisation du service téléphonique est très importante. L'opérateur a construit un immense réseau de télécommunication, mais à cause de la mauvaise configuration du système, les utilisateurs ne sont pas satisfaits avec les services, les investissements n'ont pas été remis. Donc les entreprises comme IBM, Huawei, et d'autres fournisseurs du équipement essaient de trouver la meilleure solution.

Maintenant, il y a beaucoup des gens qui travaillent dans ce domaine, nous avons trouvé beaucoup d'articles sur l'optimisation du réseau télécommunication, mais les articles sont basés sur réseau 3G ou 2G.

Il y a trois techniques qui sont beaucoup utilisées :

1. la Technique KQI ;
2. la Technique QoE ;
3. la Technique qui étudie les comportements de l'utilisateur.

4.1 la Technique KQI

:

La technique le plus souvent utilisée s'appelle 'KQI' (Key Quality Indicator) [14](#), cette méthode a été beaucoup utilisée. Et cette technique peut généralement divisé en deux étapes. D'abord, nous devons calculer le score de KPI, pour calculer le KPI en premier, il faut analyser le processus d'un service et choisir les indicateurs de performance. Ensuite, nous pouvons calculer le score d'un processus en utilisant une équation linéaire, le poids de

chaque attribut change selon le service, par exemple, pour le service SMS, le délai porte peu d'importance, mais le délai du service est important pour le service HTTP. À la fin, nous pouvons calculer le KQI avec les KPI [2]. Mais les poids sont définis par les experts, et les valeurs peuvent être fausses ou pas précises. Et par fois le score est bon mais l'expérience de l'utilisateur n'est pas bonne.

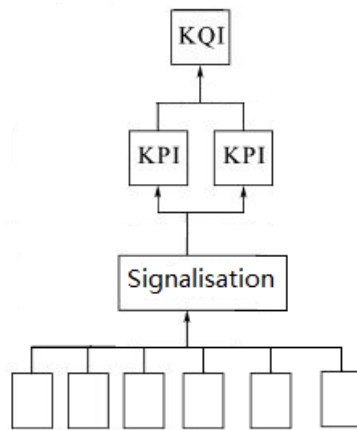


FIGURE 14 – KQI

4.2 la Technique QoE

:

KPI est un des indicateurs de qualité axés sur les performances du réseau, mais il ne reflète pas directement l'expérience de la qualité de service de l'utilisateur, parce que les expériences de l'utilisateur sont difficile à mesurer. Donc la technique QoE a été inventé. QoE définit la performance de la qualité de service et l'expérience de l'utilisateur de l'ensemble du réseau à partir de l'utilisateur.

Les utilisateurs ont nombreuses exigences pour les services téléphoniques, Elles peuvent être résumées en deux aspects : la fiabilité et le confort. La fiabilité fait référence à l'activité de l'accessibilité, la disponibilité et la durabilité. Le confort est une qualité de service, un indice de la perception directe de l'utilisateur, qui dépend de l'expérience de l'utilisateur[3]. Les relations entre QoE et QoS KPI sont : la fiabilité du service 3, le confort du service4.

TABLE 3 – Fiabilité du service

KQI	QoE
Accessibilité	Taux de succès
Disponibilité	Temps d'accès aux services
Durabilité	La durée de l'accès des services

TABLE 4 – Confort du service

KQI	QoE
	Taux de perte de paquets de couche d'application
	Le débit moyen
La qualité du service de transmission	Stabilité de la transmission
	Le bout en bout délai moyen
	Gigue
Le persistant de la connexion de service	La vitesse et la difficulté du service d'assistance

Maintenant, la technique de QoE a été beaucoup utilisée pour le service vocal. Et à cause de la complexité du service de données, il n'y a pas un standard de QoE pour le service de données.

En utilisant la technique KQI et QoE, nous pouvons mesurer la qualité du service, les résultats peuvent aider les opérateurs trouver les services de mauvaise qualité, les opérateur peuvent améliorer les services selon le résultat, finalement améliorer la notation de l'utilisateur.

Le résultat de KQI dépend seulement des performances du réseau, donc nous avons besoin des informations et des performances de réseau. Et la technologie QoE a besoin du résultat de KQI et de feed-back de l'utilisateur, le feed-back peut être obtenu par l'enquête ou les plaintes des utilisateurs et par les mesures directs.

4.3 la Technique qui étudie les comportements de l'utilisateur

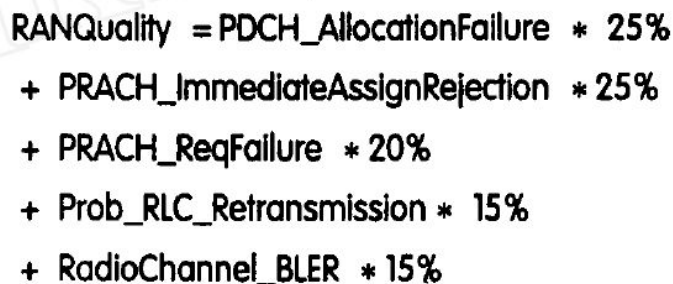
Aussi il y a un groupe qui utilise les comportements de l'utilisateur pour définir la qualité du service[4]. Le groupe utilise cette méthode dans le service vocal, il cherche la situation où l'utilisateur accroche et ré-appel le même personne. À la fin, cette méthode aide l'opérateur corriger le paramètre d'erreur.

Selon l'article, cette méthode peut aider l'opérateur trouver les défauts du système, mais il a nombreuses restrictions, par exemple, nous ne pouvons pas utiliser cette technologie dans le service de SMS, etc.

5 Le présentation de notre solution

La méthode qui utilise les comportements de l'utilisateur est intéressante, mais nous trouvons qu'elle peut utiliser seulement dans le service vocal, nous n'avons pas trouvé les règles similaires dans l'autre service. D'ailleurs, le réseau LTE ne support pas le service vocal, donc il n'existe pas d'optimisation du service vocal dans le réseau LTE et nous n'avons pas de données. Donc nous ne pouvons pas utiliser cette méthode.

La technique QoE et KQI sont beaucoup utilisés, mais d'abord, pour la méthode QoE, nous avons besoin des réponses des utilisateurs, mais nous n'avons pas assez de temps, et des raisons financières, le CMCC ne peut pas nous fournir ces données. Et l'équation qu'on utilise pour calculer KPI n'est pas convaincante, voici une exemple d'un équation pour calculer la disponibilité du réseau pour le service SMS dans réseau 3G¹⁵.



The figure shows a mathematical equation for calculating RANQuality. It is presented as a list of terms, each followed by a multiplication sign and a percentage. The terms are: PDCH_AllocationFailure (25%), PRACH_ImmediateAssignRejection (25%), PRACH_ReqFailure (20%), Prob_RLC_Retransmission (15%), and RadioChannel_BLER (15%). The entire equation is enclosed in a light blue rounded rectangular box.

$$\begin{aligned} \text{RANQuality} = & \text{PDCH_AllocationFailure} * 25\% \\ & + \text{PRACH_ImmediateAssignRejection} * 25\% \\ & + \text{PRACH_ReqFailure} * 20\% \\ & + \text{Prob_RLC_Retransmission} * 15\% \\ & + \text{RadioChannel_BLER} * 15\% \end{aligned}$$

FIGURE 15 – Un exemple d'un équation pour calculer la disponibilité

Le poids du chaque attributs est défini par les experts, Mais l'utilisateur n'est pas satisfait du service. Nous croyons que l'erreur a été causée par l'in-

exacte équation, et nous pensons que les algorithmes de classification peuvent aider à améliorer le résultat, mais très vite nous avons trouvé que le CMCC ne peut pas nous fournir ce type de données. Sans la connaissance à priori, nous ne pouvons pas utiliser ces algorithmes. Nous avons aussi pensé à utiliser le externalisation ouverte (crowdsourcing), à notre avis si le CMCC peut lancer un projet de externalisation ouverte. Si le CMCC peut encourager ses utilisateurs donner les notes aux services pour obtenir des crédits, nous pouvons obtenir la connaissance à priori, et à l'aide de ces données, nous pouvons trouver une équation peut-être mieux que les équations écrivent par les experts. Mais bien sur, le CMCC n'a pas accepté cette idée, parce que cette méthode peut coûter cher, et peut-être il n'y a pas de revenu direct. Et l'entreprise ne fait pas de l'investissement sans retour. Donc nous n'avons pas de connaissance a priori.

Finalement, nous avons décidé d'utiliser la technique de l'apprentissage non supervisé, il contient la notion de Réseau de neurones et l'algorithme de clustering etc. Nous avons choisi l'algorithme de clustering. Et nous avons utilisé la technique de l'arbre couvrant de poids minimal en anglais Minimum spanning tree (MST) et la méthode Règles d'association et PCA.

5.1 Clustering

L'algorithme de clustering est une des méthodes de classification non supervisée. Il est beaucoup utilisé quand la donnée n'a pas de connaissance a priori.

C'est une méthode statistique d'analyse des données. Elle divise un ensemble de données en différents groupes, les données de chaque groupe est mathématiquement plus proche que les données de l'autre groupe, et nous supposons que les données dans le même partition ont des caractéristiques similaires.

Il existe de multiples méthodes de regroupement des données, parmi lesquelles :

- Classification basées sur la densité ;
- Classification hiérarchique ;
- Classification par partitionnement ;
- Classification par grille ;
- Classification basées sur des modèles.

Les étapes de cette algorithme est :

- Choisir k points qui représentent la position moyenne des k partitions initiales (au hasard) ;

- Répéter les étapes suivantes jusqu'à convergence :
 1. assigner chaque observation à la partition la plus proche
 2. mettre à jour la moyenne de chaque cluster

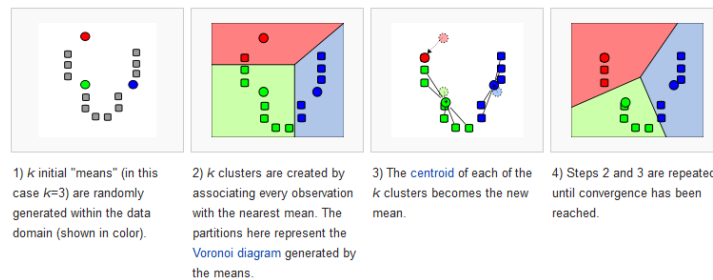


FIGURE 16 – étapes de l'algorithme K-Means

Nous avons utilisé l'algorithme K-Means et CLARA (Clustering Large Applications). Ils sont deux techniques de Classification par partitionnement.

Le but de cet algorithme est de diviser des données en K partitions (clusters) dans lesquelles les données appartiennent à la partition avec la moyenne la plus proche.

5.1.1 La distance

Quand on regroupe un ensemble de données, on calcule la distance entre chaque échantillon, la distance entre les échantillons dans un même groupe est plus petite que les échantillons dans l'autre groupe. On suppose que pour les deux échantillons les plus similaires, plus la distance est petite.

Il y a plusieurs moyens de calculer la similarité. On utilise la technique qui est basée sur la distance. Les techniques utilisées le plus souvent sont : *distance euclidienne*, *distance de Hamming*, *distance Manhattan*, etc. Aussi il existe des méthodes qui mesurent la similarité, y compris : *coefficient de Jaccard*, *cosinus similarité*, etc. Nous décidons de utiliser la distance euclidienne pour calculer la similarité.

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

n est la dimension, x_k et y_k sont le k -ème attributs de échantillon x et y .

5.1.2 La validité de clustering

Cette méthode a été utilisée dans beaucoup de différents domaines. Mais la qualité du résultat dépend du nombre de clusters ou de la valeur de K, différents paramètres peuvent mener à des résultats très différents.

En 1974, Mme BEZDEK a proposé cette question de d'évaluation, et donné la première fonction pour évaluer le résultat du clustering : **Partition Coefficient**. Par la suite, de nombreux chercheurs ont proposé une variété de fonctions pour évaluer le résultat, et évaluer les qualités et le champ d'application de ses fonctions.

Par mesure de l'efficacité entre les partitions et inter-partition, on peut évaluer le clustering. Le résultat de clustering idéal devrait être l'obtention de la distance minimale dans les partitions, et la distance maximale entre les partitions. C'est à dire avec le plus petit cohésion dans la partition, et le plus élevé degré de la séparation entre les partitions. La cohésion mesure le degré de proximité dans le groupe, et la séparation mesure le degré de dissimilarité entre les groupes. La cohésion et la séparation peuvent être calculée par les équations suivantes.

$$cohson(C_i) = \sum_{x \in 1} proximity(x, c_i)$$

$$sparation(C_i, C_j) = proximity(c_i, c_j)$$

Dans les formules, le c est le centre de gravité d'une partition, $proximity(a, b)$ est la mesure de proximité entre la partition a et b.

Le degré de la cohésion de clusters et celui séparation sont souvent utilisés comme une mesures principales pour évaluer le résultat de clustering. Un bon regroupement devrait être à la fois un petit degré de la cohésion et un grand degré de la séparation.

Le degré de la cohésion et celui de la séparation de la partition ne sont pas indépendants, la somme des deux est une constante, on suppose que quand on a le minimal degré de la cohésion, on peut avoir le maximal degré de la séparation. Évidement, nous devons utiliser à la fois le degré de la cohésion et celui séparation pour mesurer la qualité du regroupement.

5.1.3 La silhouette Coefficient

Kaufman a proposé la silhouette coefficient en 2010, cette méthode utilise à la fois les deux degrés.

1. la silhouette coefficient d'un échantillon :

Pour un échantillon d_i , en supposant qu'il est dans groupe A, la silhouette coefficient peut être calculée par cette formule :

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

a_i est la dissemblance moyenne de la donnée i avec toutes les autres données dans le même cluster (plus la valeur est petite, meilleur est le regroupement). b_i est la différence moyenne de i à tout autre groupe i n'est pas un membre.

Qui peut être écrite comme :

$$s_i = \begin{cases} 1 - \frac{a_i}{b_i}, & \text{if } a_i < b_i \\ 0, & \text{if } a_i = b_i \\ \frac{b_i}{a_i} - 1, & \text{if } a_i > b_i \end{cases}$$

D'après la définition ci-dessus, il est clair que la valeur de s_i est entre 1 et -1

$$-1 < s_i < 1$$

Pour que s_i soit proche de 1, nous demandons que le a_i est plus petit que b_i . Comme a_i mesure la dissemblance de i et son propre groupe, une petite valeur signifie qu'il est bien adapté. En outre, un grand b_i implique que i est mal adapté à son groupe voisin. Ainsi, un s_i près de 1 signifie que la donnée est concentrée de manière appropriée. Si la valeur de s_i est proche de -1, selon la même logique, nous voyons que la donnée i serait plus appropriée si elle a été regroupée dans son groupe voisin.

2. la silhouette coefficient moyenne :

Pour le résultat d'un regroupement, le silhouette coefficient égale à :

$$s_k = \frac{1}{n} \sum_{i=1}^n s_i$$

n est le nombre de données, k est le nombre de partitions, s_k est la moyenne de la silhouette coefficient. Nous pouvons utiliser s_k pour

évaluer la qualité du clustering.

Le s_i est la moyenne sur l'ensemble des données d'un cluster, il est une mesure de degré de cohésion de toutes les données du cluster. Ainsi, la moyenne de s_i de l'ensemble des données (s_k) est une mesure de la façon appropriée des données qui ont été regroupées. S'il y a trop peu de clusters, cela peut se produire lorsque un mauvais choix de k est utilisé dans l'algorithme K-Means, la silhouette Coefficient de certains groupes sera beaucoup plus petit que les autres. Donc la plot de silhouette Coefficient et la valeur moyenne de silhouette peuvent être utilisées pour déterminer le nombre de clusters (la valeur de k) optimaux pour le ensemble de données.

Après avoir trouvé les paramètres optimaux, nous pouvons utiliser l'algorithme clustering, et après avoir étudié et comparé les caractéristiques de chaque partition, nous pouvons définir la quelle partition qui représentent la mauvaise qualité de service. Le résultat peut nous aider à classifier le service.

5.2 Règles d'association

La règle d'association est une méthode populaire, elle étudie le données d'une manière approfondie. Le but est de découvrir des relations ayant un intérêt pour le statisticien entre les variables. En se basant sur le concept de relations fortes, Rakesh Agrawal et son équipe présentent des règles d'association dont le but est de découvrir des similitudes entre des produits dans des données saisies sur une grande échelle dans les systèmes informatiques des points de ventes des chaînes de supermarchés. Par exemple, une règle découvre que si un homme achète les serviettes de bébé, il est susceptible de d'acheter les bières. Une telle information peut être utile quand on veut prendre des décisions marketing.

Les règles d'association sont employées aujourd'hui dans plusieurs domaines, incluant : la fouille du web, la détection d'intrusion et la bio-informatique. Dans ces domaines, ils utilisent les données booléennes pour trouver les règles utiles. Mais dans le domaine télécommunication, les données de signalisation sont les données numériques. Donc nous ne pouvons pas directement utiliser la règles d'association. Par contre, nous pouvons convertir les données de type numérique en données de caractère en divisant les données dans plusieurs partitions [6]. Par exemple, nous avons des données numériques entre 0 et 100, et on divise les données en 10 partitions, donc les chiffres entre 0 et 10 peuvent présenter par $0 < x < 10$. En utilisant

cette technique, nous pouvons utiliser la règle d'association pour trouver les relations dans les données.

Il y a plusieurs méthodes pour diviser les attributs numériques : par catégorisation, l'analyse typologique, par analyse de l'histogramme, l'analyse basé sur l'entropie de discret, partition naturelle et ainsi de suite.

Nous décidons de utiliser le K-Means d'abord. Après avoir analysé le résultat, nous pouvons utiliser l'algorithme règles d'associations à l'aide du résultat de K-Means.

5.3 K-Means et l'arbre couvrant de poids minimal

D'après les techniques précédantes, nous avons utilisé l'arbre couvrant de poids minimal avec K-Means.

En théorie des graphes, étant donné qu'un graphe non orienté connexe et que ses arêtes sont pondérées, un arbre couvrant de poids minimal de ce graphe est un arbre couvrant (sous-ensemble qui est un arbre et qui connecte tous les sommets ensemble) dont la somme des poids des arêtes est minimale¹⁷. L'arbre couvrant de poids minimal est aussi connu sous certains autres noms, tel qu'arbre couvrant minimum ou encore arbre sous-tendant minimum. L'algorithme de Prim et l'algorithme de Kruskal sont deux méthodes classique, qui sont tous les deux l'algorithme glouton.

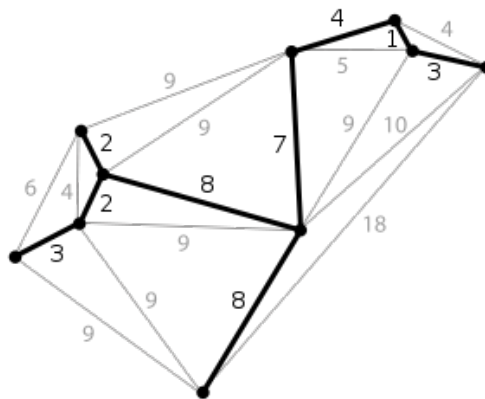


FIGURE 17 – MST

En coupant $C - 1$ arêtes la plus grande, nous pouvons grouper les données en C parties¹⁸.

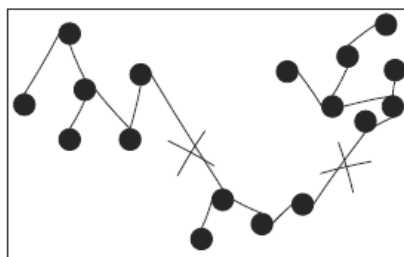


FIGURE 18 – MST clustering

Dans notre projet, nous avons choisi l'algorithme Prim et K-Means pour faire le clustering.

La complexité d'algorithme Prim est : $O((V + E) + \lg(V)) = O(E \lg(V))$, V est l'ensemble de vertex, E est l'ensemble de arêtes.

La complexité d'algorithme K-Means est : $O(nkt)$, n est la quantité de données, k est le nombre de partitions, t est le nombre d'itérations.

5.3.1 L'algorithme KmMST

Dans l'article [7], nous avons trouvé une algorithme qui utilise K-Means et la MST.

Parce que K-Means peut trouver des partitions sur forme de sphéricité, et la valeur de K n'est pas grande, donc le résultat est influencé par les données de bruits, donc le résultat de K-Means est par fois peu satisfaisant.

D'après l'article, nous pouvons choisir une grande valeur pour K , plus utiliser la technique MST pour combiner les partition en C groupes en coupant $C-1$ arête. La performance est meilleure que celui de K-Means [19](#).

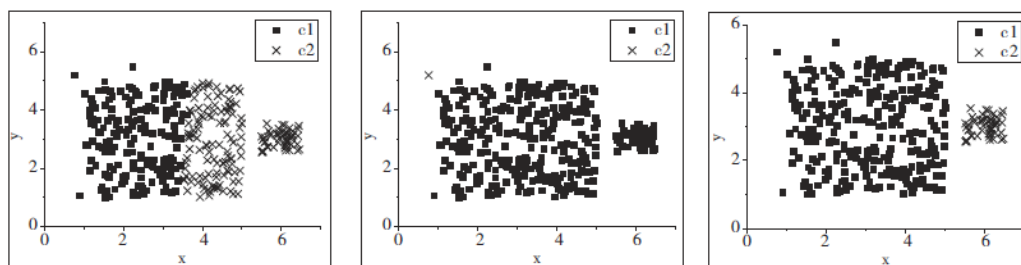


FIGURE 19 – La performance de K-Means, MST et KmMST

L'algorithme KmMST peut être décrit comme suit :

1. calculer la valeur de K , $k = \lfloor n^r \rfloor$ ($r \in [0, 1]$) par défaut la valeur de r égale à 0,5 ;
2. utiliser l'algorithme K-Means, trouver K partitions ;
3. calculer la distance entre chaque centre de la partition ;
4. utiliser l'algorithme Prim pour créer l'arbre couvrant de poids minimal ;
5. couper le $C - 1$ plus bords qui ont la plus grande distance, le C sous-graphe est le résultat de clustering.

Pour utiliser cet algorithme, nous avons besoin de la quantité n de données, du coefficient de r et de la valeur de C . En utilisant cette méthode nous pouvons trouver les partitions.

6 La mise en œuvre

6.1 Le logiciel utilisé

L'objectif de notre projet est de trouver une méthode qui peut s'implanter dans les serveurs du CMCC, et aider le CMCC à améliorer la qualité du réseau. D'après le directeur du R&D département du CMCC, au total, CMCC a 20Pb de données stockées dans sa base de données, donc le logiciel doit être capable d'exécuter une grande quantité de données. En plus, au lieu d'utiliser directement les logiciels comme SQL (qui n'est pas très efficace si on a beaucoup de données stockées dans différents serveurs), le CMCC utilise 'Hadoop' pour stocker et gérer ses données.

Donc le logiciel que nous utilisons doit être capable d'exécuter une grande quantité de données et peut travailler avec Hadoop.

Finalement nous décidons d'utiliser le langage R. Les avantages de R sont :

- R est un langage et un environnement pour le calcul statistique et les graphiques ;
- R offre une grande variété de statistiques (modélisation linéaire et non linéaire, classification, clustering, etc.) et des graphiques techniques, et il est très extensible ;
- R est facile à utiliser ;

- R est un logiciel libre, il compile et fonctionne sur une grande variété de plates-formes UNIX et les systèmes similaires (y compris FreeBSD et Linux), Windows et Mac OS ;
- en utilisant les packages fournis par 'Revolution Analytics', nous pouvons utiliser Hadoop en R.

Nous utilisons Rstudio comme notre environnement de programmation. C'est une interface d'utilisateurs puissante et productive pour R.

6.2 Introduction des données

Après quelque semaines de discussions avec les employés de différents départements de CMCC, ils nous ont fourni deux versions de données, et leurs spécifications du format[1]. Nous avons trouvé que le CMCC n'a pas d'accès direct aux données, et le fournisseur d'équipement a modifié la spécification fournie par le CMCC, et il y a des erreurs dans les données fournies par le fournisseur d'équipement.

Ils nous ont envoyé 11 dossiers, chaque dossier correspond à un service. les services sont 'rtsp', 'dns', 'mail', 'ftp', 'http-wap', 'mms', 'p2p', 'realtimecom', 'VoIP' et les données de signalisation entre E-UTRAN et MME 'S1AP-NAS'.

Et nous avons trouvé que pour les services comme 'VoIP' et 'RTSP', ils ont très peu de données 5. Donc nous avons décidé d'utiliser le donnée du service 'HTTP'.

L'interface	Nombre de ligne
S1-AP	240
RTSP	35
DNS	272562
Maill	44
FTP	71
HTTP-WAP	50854
MMS	193
P2P	515
Realtimecom	2082
S1U	89759
VoIP	28

TABLE 5 – les dossiers de données

Le dossier du service HTTP a 18,4Mbit , il y a 50854 lignes, tous les

données sont collectées par les capteurs placés entre les Service-Gateway et les eNodeB. Le capteur enregistre une ligne de données quand un processus est fini. Chaque ligne a 76 attributs [20](#).

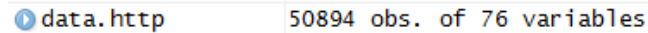


FIGURE 20 – les données du service HTTP

Il contient des informations d'un UE (IMEI, IMSI, etc), les trafics de la liaison montante et la liaison descendante et le temps, l'adresse IP de UE, eNodeB et S-GW, le port de l'eNodeB et le S-GW, le délai du service, le site web, cookie, et aussi le temps de commencer et le temps d'arrêter. Les données sont collectées dans 20.92 minutes [21](#).

```
StartTimes <-as.POSIXlt(as.numeric(substr(data_HTTP$StartT,1,10)), "UTC", origin="1970-01-01")
EndTimes   <-as.POSIXlt(as.numeric(substr(data_HTTP$StopT,1,10)), "UTC", origin="1970-01-01")
RecodeTime <-data.frame(StartTimes,EndTimes)
attach(RecodeTime)
```

```
cat("Data recoded in:",max(EndTimes)-min(StartTimes)," minits")
```

```
## Data recoded in: 20.92 minits
```

FIGURE 21 – Les données sont collectent dans 20.92 minutes

6.3 Pré-traitement de données

En analysant des données, nous avons trouvé des erreurs de données, et le fournisseur nous a confirmé que ces sont les défauts de leur système 4G. Pour les attributs 'IMSI', 'IMEI', 'MSISDN', 90 % de lignes sont vides, pour ceux qui ne sont pas vides, les contenus sont illisibles, et peuvent provoquer des erreurs de lecture [22](#). Et nous avons trouvé que les contenus de certains lignes sont bizarres.

IMSI	IMEI	MSISDN	M.TMSI	IpType
			0 1	175453935
hw2		hQp p	1	176899993
			H 1	176918718

FIGURE 22 – erreur du codage BCD

Dans ce processus, le 'Down Link Online Time' égale à 0 ms , mais il a téléchargé 746 bits, c'est clairement une erreur dd données.

BigType	SubType	L4	ServerIP	ServerPort	UpTraffic	DownTraffic	UpTime	DownTime
15	5017	0	3719544451	80	595	746	500	0

FIGURE 23 – Erreur de la donnée

Nous avons décidé de ne pas utiliser les données avec ce type d'erreurs, à la fin, en supprimant ses données, il nous reste 37865 lignes (50894 lignes en origine, 13029 lignes ont été supprimé) 24.

data_DisHttp	50894 obs. of 76 variables
data_HTTP	37865 obs. of 76 variables

FIGURE 24 – Pré-traitement des données

Entre ces 76 attributs, une grande partie de ces informations sont inutiles, et pour certains attributs les contenus égalent tous à 0. Finalement, nous avons trouvé 11 attributs. ils sont :

Signalisation	Signalisation	KPI
trafic en liaison montante	le temps en ligne	vitesse
trafic en liaison descendante	le temps en ligne	vitesse
	Http First Response Time	délai
	Http Last Packet Time	délai
	Http Last Ack Time	délai
Packet Num en liaison montante	retransmission de paquets Num en liaison montante	taux de retransmission
Packet Num en liaison descendante	retransmission de paquets Num en liaison descendante	taux de retransmission

Mais nous avons trouvé que dans certaines lignes, le taux de retransmission est trop grands (plus grand que 100%). par exemple, dans un processus, l'UE a téléchargé 17 paquets IP, et il y a 7448 paquets qui sont du désordre, et il a ré-téléchargé 6384 paquets [25](#). Les données ne sont pas corrects, donc nous ne pouvons pas utiliser ses donnée pour calculer le taux de retransmission.

UpPac	DownPac	UpDisPac	DownDisPac	UpRePac	DownRePac
6	5	0	0	0	0
10	19	0	0	0	0
5	4	380	0	0	0
8	17	0	7448	40	6384
3	6	0	2128	0	0

FIGURE 25 – Défaut de la système

Finalement nous avons décidé d'utiliser ces 5 attributs (la vitesse et le délai) pour mesurer la qualité du service.

6.4 Les caractéristiques de données

```
> min(upAvBand)      > max(upAvBand)      > sd(upAvBand)
[1] 0.06616667        [1] 170.16          [1] 1.631399
> min(downAvBand)    > max(downAvBand)    > sd(downAvBand)
[1] 0                  [1] 1316             [1] 24.36523
> min(firstRespondTime) > max(firstRespondTime) > sd(firstRespondTime)
[1] 0                  [1] 150695           [1] 4475.789
> min(lastPacketTime) > max(lastPacketTime) > sd(lastPacketTime)
[1] 0                  [1] 92081            [1] 2673.268
> min(lastAckTime)    > max(lastAckTime)    > sd(lastAckTime)
[1] 0                  [1] 1220150          [1] 8989.971
```

FIGURE 26 – La valeur maximal, valeur minimal et l'écart type

```
> mean(upAvBand)
[1] 0.85507
> mean(downAvBand)
[1] 4.735396
> mean(firstRespondTime)
[1] 594.1533
> mean(lastPacketTime)
[1] 392.4637
> mean(lastAckTime)
[1] 1231.584
```

FIGURE 27 – La valeur moyenne

Nous avons trouvé que l'écart type pour les trois attributs de délais est très grand. Les valeurs varient de $0ms$ à $1220150ms$. Par contre le changement de la vitesse de la liaison montante et descendante n'est pas très grand.

```
> data_PrepAR[,3]
[1] 6102 43 0 0 0 249 424 0 0 52 0 10 67 0 0 0 0 30125 57
[20] 23 0 0 0 0 508 0 0 0 0 0 0 0 1249 0 0 0 0 0
[39] 0 195 0 0 33 0 0 0 0 0 0 0 0 1104 0 0 0 0 0
[58] 37 0 29 0 66 0 0 0 0 0 0 0 39 0 0 0 0 0 33
[77] 0 0 236 0 0 0 0 101 0 0 742 0 0 138 0 9 1509 34 669
[96] 0 0 0 0 43 0 14 0 35 0 0 42 0 0 0 0 0 0 54
[115] 0 0 144 229 69 0 0 0 10 0 0 0 0 100 42 516 0 595
[134] 0 0 0 0 360 0 0 0 0 791 0 0 0 0 0 182 0 1271
[153] 0 0 0 0 49572 0 6159 0 25 0 0 0 0 461 1233 0 1431 0
[172] 242 0 0 36 0 76049 31 0 0 24 0 0 1347 128 0 0 0 0
[191] 0 0 0 39 59 0 120 102 1109 0 0 663 0 0 26 0 0 1083
[210] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[229] 0 0 0 116 0 0 0 0 37 0 0 0 1797 0 170 0 0 0
[248] 0 26 0 360 0 0 0 10 151 0 0 207 0 0 0 0 30249 0
```

FIGURE 28 – Le temps de réponse du serveur

En analysant les données, nous avons trouvé que les données ne sont pas normales, et l'ingénieur du fournisseur de l'équipement nous ont confirmé que le système de collecte d'informations de signalisation a des défauts.

6.5 Le K-Means et Le Règle d'association

6.5.1 La valeur optimale de k

Pour trouver le k optimal pour nos données, nous avons utilisé la technique 'silhouette Coefficient' et 'somme d'erreurs carrées'.

Nous avons mesuré le résultat de ces deux techniques quand la valeur de k augmente de 1 à 15. En tenant compte des caractéristiques de l'algorithme, pour chaque valeur de k , nous avons répété 50 fois en vue de calculer la valeur moyenne pour éliminer les erreurs. En suite, nous avons étudié le résultat pour trouver la valeur optimale.

La somme d'erreurs carrées

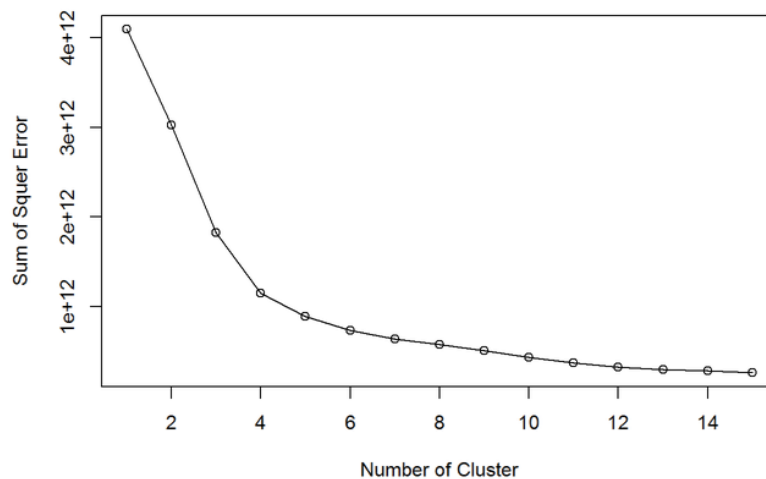


FIGURE 29 – somme d'erreurs carrées

Nous avons trouvé que la somme d'erreurs carrées diminue quand k augmente, mais il est difficile de choisir la valeur optimale de k avec cette image.

La silhouette Coefficient

Nous utilisons la même technique pour calculer et visualiser la valeur de la silhouette Coefficient en fonction de la valeur de k .

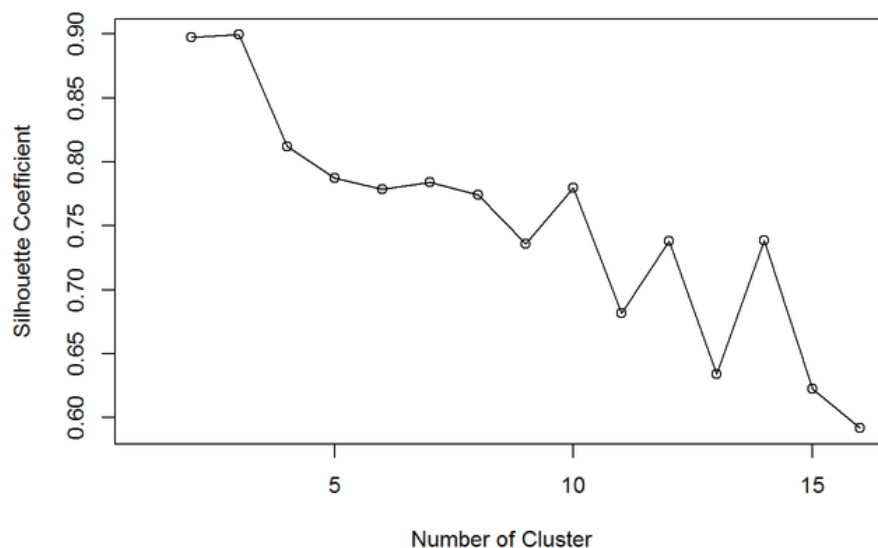


FIGURE 30 – la silhouette Coefficient quand k varie de 2 à 15

L'image de la silhouette Coefficient nous montre que comme quand $k = 3$ la valeur de la silhouette Coefficient est la plus grandes. Donc nous soupçonnons que quand $k = 3$, nous pouvons trouver la meilleure partition.

6.5.2 Le clustering

Après avoir classé le données en 3 groupes en utilisant l'algorithme CLARA, nous avons trouvé que la majorité de données (94% de données) sont dans le premier groupe, un groupe de 4,4%, le troisième groupe a 1,6% de données.



FIGURE 31 – Le nombre des trois clusters

En utilisant la fonction 'plot3d' fournie par le package 'rgl', nous pouvons visualiser les données en trois dimensions, et nous pouvons visualiser le résultat de clustering.

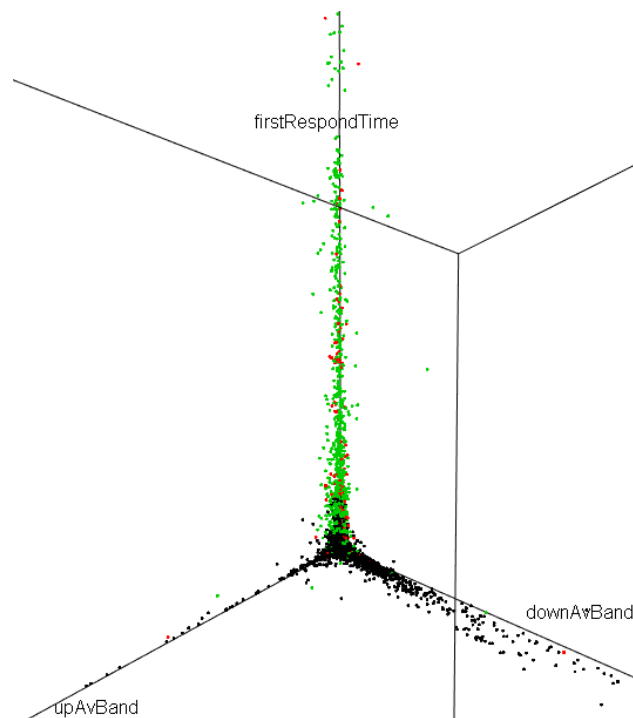


FIGURE 32 – up link Average Band, down link Average Band, first Respond Time.

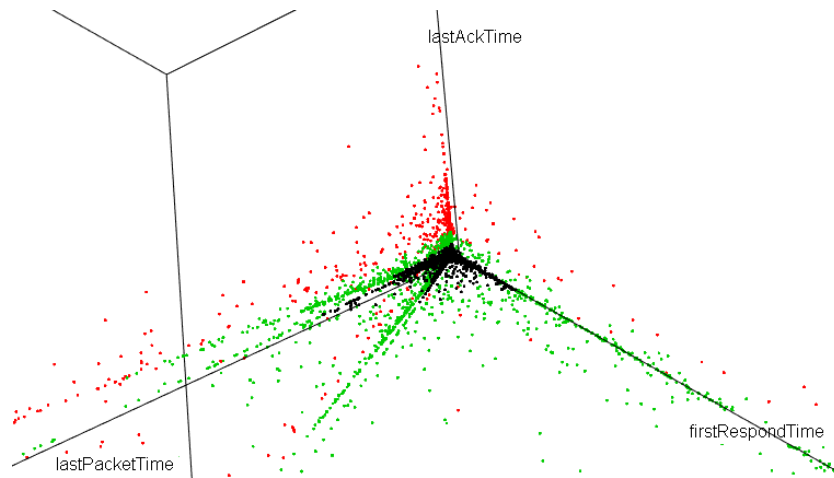


FIGURE 33 – first Respond Time, last Packet Time, last Ack Time.

Nous avons trouvé que les points noirs sont les données de 'cluster 1', les points de 'cluster 2' sont en rouge, les points de 'cluster 3' sont en vert.

Après le regroupement, nous pouvons utiliser l'algorithme 'k plus proches voisins' pour déterminer le groupe dont les nouvelles données font partie.

6.5.3 La Règle d'association

En utilisant le résultat de clustering, nous pouvons transformer les données numériques aux données caractéristiques.

Nous y parvenons en quatre étapes

1. D'abord, nous trouvons les valeurs minimale et maximale de chaque attribut³⁴;
2. Ensuite, nous trouvons les intervalles de chaque attribut³⁵;
3. Puis, nous transformons les données numériques aux données caractéristiques³⁶;
4. À la fin, nous utilisons l'algorithme Apriori pour trouver des règles³⁸.

	row.names	V1	V2	V3	V4	V5
1	minV	0.06616667	0.0000	0	0	0
2	maxV	170.16000000	1316.0000	8590	8610	7884
3	minV	0.10936691	0.1290	0	3	17902
4	maxV	40.90133333	1004.1145	93496	83559	1220150
5	minV	0.12884211	0.0800	0	4	0
6	maxV	30.15066667	243.9792	150695	92081	54999

FIGURE 34 – trouver les valeurs minimal et maximal de chaque attribut

	row.names	V1	V2	V3	V4	V5
1	seuil	0.08776679	0.0400	4295.0	1.5	3942.0
2	seuil	0.11910451	0.1045	51043.0	3.5	12893.0
3	seuil	15.13975439	122.0541	122095.5	4307.0	36450.5
4	seuil	35.52600000	624.0468	0.0	46084.5	637574.5
5	seuil	105.53066667	1160.0573	0.0	87820.0	0.0

FIGURE 35 – définir les intervalles

	row.names	upAvBand	downAvBand	firstRespondTime	lastPacketTime	lastAckTime
1	1	3	3	2	3	1
2	2	3	4	2	3	5
3	3	3	4	1	1	1
4	4	2	4	1	1	1
5	5	3	4	1	1	1
6	6	3	4	1	2	1
7	7	3	4	1	3	1
8	8	3	3	1	1	1
9	10	3	3	1	1	1
10	11	3	3	4	5	1
11	12	3	3	1	1	1

FIGURE 36 – transformer les données numériques aux données caractéristique

En changeant les paramètres de l'algorithme, nous pouvons filtrer le résultat. Par exemple, nous cherchons les règles qui ont le support supérieur à 0,5 et la confiance supérieure à 0,8, et aussi la valeur de 'lastAckTime' et 'upAvBand' sont dans la première intervalle³⁷. Et nous avons trouvé 7 règles à la fin.³⁸

```
system.time(aRule<-apriori(trans,parameter = list(minlen=2, supp=0.5, conf=0.8)
, appearance = list(rhs=c("lastAckTime=1", "upAvBand=1")
, default="lhs"),control = list(verbose=F)))
```

FIGURE 37 – L’algorithme règle d’association et le filtrage des données

```
> summary(aRule)
set of 7 rules

rule length distribution (lhs + rhs):sizes
2 3 4
3 3 1

      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2.000  2.000   3.000   2.714   3.000   4.000

summary of quality measures:
      support      confidence      lift
Min.   :0.6551  Min.   :0.8929  Min.   :1.003
1st Qu.:0.6823  1st Qu.:0.9239  1st Qu.:1.038
Median :0.7096  Median :1.0000  Median :1.123
Mean   :0.7339  Mean   :0.9630  Mean   :1.081
3rd Qu.:0.7895  3rd Qu.:1.0000  3rd Qu.:1.123
Max.   :0.8293  Max.   :1.0000  Max.   :1.123

mining info:
 data ntransactions support confidence
trans      37865      0.5      0.8
```

FIGURE 38 – Les règles trouvées par l’algorithme règle d’association

```
> inspect(aRule)
lhs      rhs      support confidence lift
1 {lastPacketTime=1} => {lastAckTime=1} 0.7095999 1.0000000 1.123091
2 {firstRespondTime=1} => {lastAckTime=1} 0.8292619 0.9230679 1.036689
3 {upAvBand=3} => {lastAckTime=1} 0.8166381 0.8928994 1.002807
4 {firstRespondTime=1, lastPacketTime=1} => {lastAckTime=1} 0.7095999 1.0000000 1.123091
5 {upAvBand=3, lastPacketTime=1} => {lastAckTime=1} 0.6550640 1.0000000 1.123091
6 {upAvBand=3, firstRespondTime=1} => {lastAckTime=1} 0.7622871 0.9248022 1.038637
7 {upAvBand=3, firstRespondTime=1, lastPacketTime=1} => {lastAckTime=1} 0.6550640 1.0000000 1.123091
```

FIGURE 39 – Les règles découvertes

Selon les règles trouvées par l’algorithme de la Règle d’association, si la valeur de ‘lastPacketTime’ est dans la deuxième intervalle 1, la valeur de ‘lastAckTime’ sera plus susceptible d’être dans l’intervalle 1, et entre toutes les données ou si la valeur de ‘firstRespondTime’ est dans la première intervalle, .82,9% d’entre eux la valeur de ‘lastAckTime’ sera dans l’intervalle 1.

En utilisant les paramètres différent, nous pouvons trouver la règle dont nous avons besoin. Mais dans notre cas, les règles trouvées ne sont pas claires.

6.6 Le KmMST

La deuxième méthode, nous utilisons l'algorithme K-Means avec MST.

Parce que l'algorithme K-Means ne fonctionne bien qu'avec les données de distribution sphérique, donc par clustering le données en beaucoup de partitions, et en utilisant la technique de MST, nous pouvons surmonter les défaut de K-Means et obtenir les bons résultats de clustering.

6.6.1 L'étape de l'algorithme KmMST

1. Regroupement des données ;

La valeur de K égale à n^r , le n égale à la quantité de données, r est dans l'intervalle de 0 et 1, par défaut, r égale à 0,5.

En premier, nous avons regroupé les données en K partition.

2. Établir la matrice de distance ;

Après, nous pouvons utiliser les valeurs du centre de chaque partition pour calculer la matrice de distance

	weight	from	to
1	6110.7599	1	2
2	6110.7599	1	3
3	6110.7562	1	4
4	6110.7605	1	5
5	5439.0350	1	6
6	48387.2799	1	7
7	5579.4425	1	8
8	6111.3178	1	9
9	5809.5372	1	10
10	10534.8457	1	11

FIGURE 40 – la matrice de distance

3. Utilises la technique de MST ;

À l'aide du package 'igraph', nous pouvons trouver l'arbre couvrant de poids minimal.

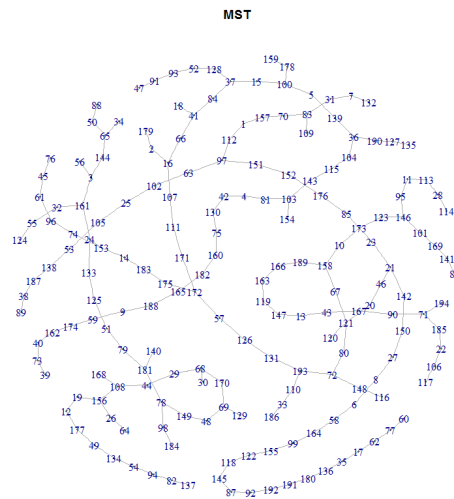


FIGURE 41 – L'arbre couvrant de poids minimal

4. Couper le arête la plus longue

Techniquement, nous pouvons trouver les partitions par couper les arêtes longue, et nous avons testé avec différentes valeurs, mais les sommets sont devenus les points isolés.

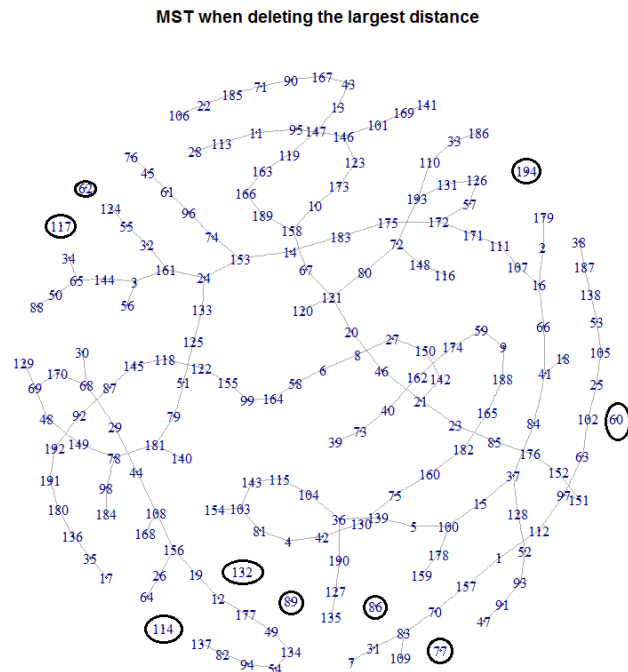


FIGURE 42 – L'arbre couvrant de poids minimal, quand $C = 10$

Conclusion

Pendant ces 4 mois de stage, nous avons lu beaucoup d'articles, nous avons préparé des dossiers et fait des rapports pour le CMCC, aussi nous avons négocié avec les gens du CMCC et les gens du fournisseur de l'équipement. Nous avons perdu beaucoup de temps en négociation en attente, mais finalement, nous avons trouvé 2 méthodes pour résoudre le problème du CMCC.

Dans le rapport, j'ai présenté les algorithmes et la mise en œuvre, mais malheureusement, les résultats ne sont pas satisfaisants, nous avons essayé différents paramètres, différentes techniques, différents logiciels, mais nous n'avons pas trouvé un résultat satisfaisant. Donc nous croyons que cela est causé du fait que la qualité de données est mauvaise. Et parce que nous avons seulement des données erronées, nous ne pouvons pas évaluer nos techniques.

Cependant, pour l'entreprise CMCC, nous avons fait une démonstration sur l'utilisation des techniques de fouille de données avec ses données, et sur les inconvénients de son système. Et aussi nous montrons que si les données ne sont pas erronées, comme il pourra trouver les informations dont il a besoin.

Pour moi, j'ai utilisé le langage R et une variété d'algorithmes pour répondre à la demande du CMCC. J'ai installé l'Hadoop dans mon ordinateur et géré l'Hadoop en Rstudio. et pendant ce stage, J'ai rencontré des bons amis dans le laboratoire, et acquis des expériences professionnelles.

Pour les membres de notre groupe, ils devront améliorer les techniques que nous avons testé et appliquer ces techniques en Hadoop en utilisant le langage R quand les nouvelles données seront arrivées.

Références

- [1] R&D DÉPARTEMENT DE CMCC. *Interface Specification of China Mobile Signaling Monitoring System(LTE Signal Collection Gateway Part)*.
- [2] Jianhua DU, Shiwen LU, Fangfeng ZHANG *Research of KQI Development Methodology in SQM*,2008.
- [3] Luning ZHAO, Zhuo SUN, Wenbo WANG *Mobile streaming QoE index system and quantify*,2012.
- [4] Rui WANG, Fei SU, Zhengdong HAN, Zilong CAI. *The Recessive Problem Mining and Optimization Research of Voice Service Based on User Behaviors*. édition, 2013.
- [5] Lianjiang ZHU, Bingxian MA, Xuequan ZHAO. *Clustering validity analysis based on silhouette coefficient*, 2010.
- [6] Hongyan WANG, Daiwen WU. *Discussion on digging algorithm of correlation rule for numerical attribute*, 2012.
- [7] Hao OUYANG, Bo CHEN, Zhenjin HUANG, Meng WANG, Zhiwen WANG. *MST Clustering Algorithm Based on K-Means*. 2014.